Consequences of local gauge symmetry in empirical tight-binding theory

Bradley A. Foreman*
Department of Physics and Institute of Nano Science and Technology,
The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

A method for incorporating electromagnetic fields into empirical tight-binding theory is derived from the principle of local gauge symmetry. Gauge invariance is shown to be incompatible with empirical tight-binding theory unless a representation exists in which the coordinate operator is diagonal. The present approach takes this basis as fundamental and uses group theory to construct symmetrized linear combinations of discrete coordinate eigenkets. This produces orthogonal atomic-like "orbitals" that may be used as a tight-binding basis. The coordinate matrix in the latter basis includes intra-atomic matrix elements between different orbitals on the same atom. Lattice gauge theory is then used to define discrete electromagnetic fields and their interaction with electrons. Local gauge symmetry is shown to impose strong restrictions limiting the range of the Hamiltonian in the coordinate basis. The theory is applied to the semiconductors Ge and Si, for which it is shown that a basis of 15 orbitals per atom provides a satisfactory description of the valence bands and the lowest conduction bands. Calculations of the dielectric function demonstrate that this model yields an accurate joint density of states, but underestimates the oscillator strength by about 20% in comparison to a nonlocal empirical pseudopotential calculation.

I. INTRODUCTION

Tight-binding theory was originally proposed as an *ab initio* technique for calculating the electronic properties of crystalline solids from atomic wave functions.[1] However, first-principles calculations based on a linear combination of atomic orbitals (LCAO) are computationally very demanding, and the tight-binding approach met with relatively little success until Slater and Koster suggested that it be used as an interpolation scheme,[2] in which the Hamiltonian matrix elements are fitted to experimental data or to band structures computed by other methods. This made it possible to describe atomic-level physics in a basis of minimal size, leading to wide-ranging applications in many areas of condensed-matter physics.[3,4,5,6,7,8] With modern computer capabilities, first-principles electronic-structure calculations are now commonplace, and *ab initio* tight-binding theories are flourishing.[6,7,8] Yet even today, the empirical theory[2] predominates (even for the fitting of first-principles calculations) because it is simple and physically intuitive.

The formalism of Slater and Koster[2] is incomplete, however, in that it contains no prescription for coupling the electronic system to external electromagnetic fields. In *ab initio* theories,[6,7,8] one can use minimal coupling (with suitable modifications for nonlocal potentials[9]) and calculate directly the necessary matrix elements of the momentum or velocity operator. In the empirical theory, these matrix elements can simply be treated as extra fitting parameters,[10,11,12] determined by fitting the dielectric function (and thus oscillator strengths) to experimental or first-principles spectra. However, even with the full use of symmetry restrictions, the number of additional parameters can be undesirably large; for example, Chang and Aspnes[11] have proposed an sp^3d^2 model for GaAs with 13 Hamiltonian parameters and 17 independent momentum parameters.

It is therefore clearly desirable to find ways of reducing or eliminating these extra fitting parameters. One possibility is to define a kinematic momentum operator (equal to mass m times velocity) by

$$\mathbf{p} = \frac{m}{i\hbar}[\mathbf{x}, H], \tag{1.1}$$

where $\mathbf{x}$ is the coordinate of the electron and H is the Hamiltonian. In a sense this merely trades one problem for another, since the coordinate matrix elements are still unknown, and the number of these allowed by symmetry is no less than the number of momentum matrix elements. However, it is physically reasonable to simplify the coordinate matrix by setting all nonlocal matrix elements to zero:

$$\langle \alpha, \mathbf{x}_i | \mathbf{x} | \alpha', \mathbf{x}_{i'} \rangle = \delta_{ii'}[\delta_{\alpha\alpha'}\mathbf{x}_i + \mathbf{x}_{\alpha\alpha'}(i)]. \tag{1.2}$$

Here $|\alpha, \mathbf{x}_i\rangle$ is the ket vector for an orthogonalized atomic orbital (Löwdin orbital[13,14]) of type α located at position $\mathbf{x}_i$. The parameter $\mathbf{x}_{\alpha\alpha'}$ is an intra-atomic matrix element coupling orbitals α and α' on the same atom.

The simplest choice of all is to set

$$\mathbf{x}_{\alpha\alpha'} \equiv 0; \tag{1.3}$$

in this model, there are no fitting parameters beyond those found in the Hamiltonian.[3,4,15,16,17,18,19,20,21] A closely related approach is the Peierls substitution,[9,22,23,24,25,26] in which the zero-field Hamiltonian matrix $\langle \alpha, \mathbf{x}_i | H | \alpha', \mathbf{x}_{i'} \rangle$ for a particle of charge e is replaced by

$$\langle \alpha, \mathbf{x}_i | H | \alpha', \mathbf{x}_{i'} \rangle \exp\left(\frac{ie}{\hbar c} \int_{\mathbf{x}_{i'}}^{\mathbf{x}_i} \mathbf{A} \cdot \mathrm{d}\mathbf{x} \right) + e\phi(\mathbf{x}_i)\delta_{ii'}\delta_{\alpha\alpha'} \tag{1.4}$$

in the presence of a vector potential $\mathbf{A}$ and scalar potential ϕ. If the path of integration in Eq. (1.4) is chosen to be a straight line,[9,23,25] then the linear term in the

Taylor series expansion of this equation is the same as the $\mathbf{A} \cdot \mathbf{p}$ coupling obtained from Eqs. (1.1)–(1.3).[23]

The total elimination of extra fitting parameters makes this model an attractive one. However, by eliminating the intra-atomic matrix elements $\mathbf{x}_{\alpha\alpha'}$, one obtains a tight-binding model that is not valid in the tight-binding limit of isolated atoms. Thus, although the model should provide a reasonable description of inter-atomic transitions between extended states, one has less confidence in its ability to describe localized states, which may be important at surfaces or interfaces. Many authors have therefore suggested augmenting the zero-parameter model by including a small number of intra-atomic matrix elements.[27,28,29,30,31,32,33] It has been shown for porous Si that, although the intra-atomic matrix elements are small in magnitude (in a Bloch-function basis[1]), the *interference* between these terms and the inter-atomic matrix elements contributes 25% of the total absorption.[32] Thus, it appears that a quantitative treatment of nanostructures may not be possible (in general) without the inclusion of intra-atomic matrix elements.

The main difficulty with such models[27,28,29,30,31,32,33] is that they are not gauge invariant.[26] As shown by the examples in Refs. 9 and 34, lack of gauge invariance can lead to gross qualitative errors in the predicted values of physical quantities. Thus, there are significant problems with both approaches considered above. The models with $\mathbf{x}_{\alpha\alpha'} = 0$ are gauge invariant, but they cannot describe intra-atomic transitions. The models with $\mathbf{x}_{\alpha\alpha'} \neq 0$ can describe intra-atomic transitions, but they are not gauge invariant.

The purpose of this paper is to demonstrate a technique for constructing tight-binding models that are gauge invariant and provide a full description of intra-atomic transitions. This is achieved by treating empirical tight-binding theory not as an approximation derived from the Schrödinger equation, but as a fundamental quantum-mechanical system in its own right. This theory is required to satisfy all of the basic principles of quantum mechanics, the most important of which (in the present context) is the principle of local gauge symmetry.[35,36,37,38,39,40] The essence of this principle is the concept that electromagnetism in quantum mechanics is the gauge-invariant manifestation of a nonintegrable (i.e., path-dependent) phase factor.[36,38]

As will be shown in Sec. II below, the reason why existing models with intra-atomic coupling[27,28,29,30,31,32,33] are not gauge invariant is that the coordinates x, y, and z do not commute.[26] Gauge invariance requires commuting coordinates, the existence of which implies the existence of a basis of coordinate eigenkets. Since empirical tight-binding theory deals with finite vector spaces, the coordinate basis is necessarily discrete. Hence, the most general gauge-invariant finite vector space is a set of discrete coordinate eigenkets. This basis may be transformed to a tight-binding basis by constructing "orbitals" from symmetrized combinations of coordinate eigenkets (using well-known techniques for symmetrizing plane waves[41,42]).

The concept of gauge symmetry on a discrete lattice is not new, having appeared many years ago as a technique for imposing a momentum cutoff in quantum chromodynamics.[40,43,44,45,46] Governale and Ungarelli[34] have recently suggested using lattice gauge techniques in empirical tight-binding theory. However, their proposal, like most applications of lattice gauge theory, is based on a simple cubic lattice. As shown below, the simple cubic lattice is unsuitable for practical tight-binding models because it can only achieve sufficient accuracy with an unreasonably large basis (i.e., a very small lattice constant). Thus, the development of efficient tight-binding models requires consideration of more general geometries.

Christ, Friedberg, and Lee[47, 48, 49] have developed a lattice gauge theory for *random* lattices, which (with some slight modifications) is sufficiently general for the present purposes. However, the complete formal machinery of quantum chromodynamics is somewhat cumbersome when one is dealing only with simple electromagnetism. Thus, for reasons of clarity, the author has chosen to present the theory in terms of a simple but elegant approach used by Dirac.[36] After a preliminary discussion of topology (i.e., how an electron is permitted to move from one lattice site to another) in Sec. III, Sec. IV presents an adaptation of Dirac's analysis[36] to the case of a discrete lattice. The outcome is a gauge-invariant formulation of electromagnetism in empirical tight-binding theory.

Although the theory derived in this way has many similarities with conventional tight-binding theory, there are significant differences as well. Not all tight-binding models can be made gauge invariant;[50] this is possible only if the basis can be constructed from symmetrized coordinate eigenkets. In addition, local gauge symmetry imposes strong restrictions on the Hamiltonian matrix, which have the effect of sharply reducing the number of allowed Hamiltonian fitting parameters. Finally, unlike previous empirical tight-binding theories, the present approach provides an explicit (discrete) wave function for the electron.

The formalism derived here is applied to two semiconductors with the diamond structure (Ge and Si) in Sec. V. For these systems, a basis of 15 orbitals per atom is shown to provide a satisfactory fit to the valence bands and the lowest conduction bands (up to about 5 eV above the valence-band maximum). These results are comparable to those obtained from a 10-orbital basis proposed recently in Ref. 51. The basis used here is 50% larger, but their model[51] cannot be made gauge invariant if intra-atomic coupling is included. Thus, it appears that some tradeoffs are necessary if gauge invariance is to be achieved.

II. COORDINATE MATRICES AND THE COORDINATE REPRESENTATION

As mentioned above, the intra-atomic coupling used in existing tight-binding models[27,28,29,30,31,32,33] leads to a lack of gauge invariance.[26] This may be seen from a simple sp^3 model for a single atom. In this case, we know from atomic physics that there are coordinate matrix elements coupling the s and p orbitals:

$$\langle s|x|p_x\rangle = \langle s|y|p_y\rangle = \langle s|z|p_z\rangle \equiv c, \tag{2.1}$$

where c is real. In the basis $\{|s\rangle, |p_x\rangle, |p_y\rangle, |p_z\rangle\}$, the matrices representing x and y are therefore

$$x = \begin{bmatrix} 0 & c & 0 & 0 \\ c & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad y = \begin{bmatrix} 0 & 0 & c & 0 \\ 0 & 0 & 0 & 0 \\ c & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2.2}$$

But this implies that x and y do not commute:

$$xy - yx = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & c^2 & 0 \\ 0 & -c^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \neq 0. \tag{2.3}$$

This means that the coordinate representation (consisting of simultaneous eigenkets of x, y, and z) does not exist. Even more important, it means that the theory cannot be gauge invariant. In a gauge transformation, the vector and scalar potentials transform as

$$\mathbf{A} \to \mathbf{A} + \nabla\Lambda, \qquad \phi \to \phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t}, \tag{2.4}$$

and the state ket $|\psi\rangle$ transforms as

$$|\psi\rangle \to U|\psi\rangle, \qquad U \equiv \exp\left[\frac{ie\Lambda(\mathbf{x},t)}{\hbar c}\right], \tag{2.5}$$

where Λ is an arbitrary function of $\mathbf{x} = (x, y, z)$ and t. If a theory is gauge invariant, all physically measurable quantities must be independent of such transformations. But the expectation value $\langle x\rangle$ is a measurable quantity, and under a gauge transformation one has

$$\langle x\rangle \to \langle U^\dagger x U\rangle, \tag{2.6}$$

where $U^\dagger x U \neq x$ if Λ depends on y or z. Hence, no theory can be gauge invariant if x, y, and z do not commute.

Is there any way of achieving gauge invariance without setting $c = 0$? Perhaps the sp^3 basis is too small, and the situation might be improved by including more orbitals $(d, f, \dots)$. However, one soon finds that for any *finite* LCAO basis, the lack of gauge invariance persists. This follows directly from the Wigner-Eckart theorem—since $\mathbf{x}$ is a vector operator, it couples states with angular momentum l to those with $l\pm1$. Hence, any finite truncation of the basis results in non-commuting coordinates.

Another possibility is to keep the same sp^3 basis, but modify the coordinate matrix. The physical justification for doing so is the fact that the orbitals used in empirical tight-binding theory are not atomic orbitals; they are *orthogonalized* atomic orbitals.[13] Therefore, they do not have the full rotational symmetry of atomic orbitals; they have only the site symmetry of the crystal structure. For example, the atoms in a diamond crystal have site symmetry T_d.[52] Therefore, the orbital that was denoted $|p_z\rangle$ above should really be written as $|\Gamma^z_{15}\rangle$, since it belongs to the Γ_{15} representation of T_d.[53]

However, the d orbital $|d_{xy}\rangle$ also transforms as $|\Gamma^z_{15}\rangle$. Thus, in the T_d group, the matrix element $b \equiv \langle\Gamma^x_{15}|y|\Gamma^z_{15}\rangle$ is allowed, and the coordinate matrices (2.2) become

$$x = \begin{bmatrix} 0 & c & 0 & 0 \\ c & 0 & 0 & 0 \\ 0 & 0 & 0 & b \\ 0 & 0 & b & 0 \end{bmatrix}, \qquad y = \begin{bmatrix} 0 & 0 & c & 0 \\ 0 & 0 & 0 & b \\ c & 0 & 0 & 0 \\ 0 & b & 0 & 0 \end{bmatrix}. \tag{2.7}$$

This yields

$$xy - yx = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & c^2 - b^2 & 0 \\ 0 & b^2 - c^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.8}$$

which is equal to zero if $b = \pm c$.

Setting $b = c$ would imply that the p and d orbitals have equal weight in the Γ_{15} states. This is not as absurd as it sounds; Boguslawski and Gorczyca[54] have shown using first-principles pseudopotential calculations that for the Γ_{15} states at the top of the valence band in GaAs, the probability of finding an electron in a cation d orbital is *greater* than that of finding it in a cation p orbital. (In AlAs, the probability ratio is greater than 2.[54]) Thus, it is not unreasonable to assume that b and c have comparable magnitudes.

If one sets $b = c$, then the coordinate operators have simultaneous eigenkets $|x', y', z'\rangle$, which are given by

$$\begin{aligned} |ccc\rangle &= \tfrac{1}{2}(|\Gamma_1\rangle + |\Gamma^x_{15}\rangle + |\Gamma^y_{15}\rangle + |\Gamma^z_{15}\rangle), \\ |c\bar{c}\bar{c}\rangle &= \tfrac{1}{2}(|\Gamma_1\rangle + |\Gamma^x_{15}\rangle - |\Gamma^y_{15}\rangle - |\Gamma^z_{15}\rangle), \\ |\bar{c}c\bar{c}\rangle &= \tfrac{1}{2}(|\Gamma_1\rangle - |\Gamma^x_{15}\rangle + |\Gamma^y_{15}\rangle - |\Gamma^z_{15}\rangle), \\ |\bar{c}\bar{c}c\rangle &= \tfrac{1}{2}(|\Gamma_1\rangle - |\Gamma^x_{15}\rangle - |\Gamma^y_{15}\rangle + |\Gamma^z_{15}\rangle). \end{aligned} \tag{2.9}$$

Note that the coordinate eigenvalues are located at the corners of a tetrahedron. In fact, the linear combinations given in (2.9) are identical to the hybrid bond orbitals used in analytical tight-binding theories,[3,4] although these are not ordinarily interpreted as exact coordinate eigenkets because the Γ_{15} states are assumed to be pure p orbitals.

The procedure outlined above is rather clumsy; one simply modifies the coordinate matrices by trial and error in an attempt to make them commute. One cannot predict in advance whether the attempt will succeed, and

in general it will not. However, the unitary transformation (2.9) may be inverted to obtain

$$\begin{aligned}
|\Gamma_1\rangle &= \tfrac{1}{2}(|ccc\rangle + |c\bar{c}\bar{c}\rangle + |\bar{c}c\bar{c}\rangle + |\bar{c}\bar{c}c\rangle),\\
|\Gamma_{15}^x\rangle &= \tfrac{1}{2}(|ccc\rangle + |c\bar{c}\bar{c}\rangle - |\bar{c}c\bar{c}\rangle - |\bar{c}\bar{c}c\rangle),\\
|\Gamma_{15}^y\rangle &= \tfrac{1}{2}(|ccc\rangle - |c\bar{c}\bar{c}\rangle + |\bar{c}c\bar{c}\rangle - |\bar{c}\bar{c}c\rangle),\\
|\Gamma_{15}^z\rangle &= \tfrac{1}{2}(|ccc\rangle - |c\bar{c}\bar{c}\rangle - |\bar{c}c\bar{c}\rangle + |\bar{c}\bar{c}c\rangle),
\end{aligned} \tag{2.10}$$

which immediately suggests a more fruitful approach. The linear combinations given in Eq. (2.10) are just what one would obtain by starting with a single coordinate eigenket (say $|ccc\rangle$) and using the symmetry operations of the tetrahedral group T_d to construct "symmetrized" orbitals[41,42] that transform according to the irreducible representations of T_d.[53]

Thus, in this alternative approach, the coordinate basis is taken as fundamental, and the tight-binding basis is merely a secondary alternative that is useful for reasons of symmetry. Since the existence of a coordinate representation is necessary for gauge invariance, no tight-binding basis can be made gauge invariant if it cannot be represented in terms of symmetrized coordinate eigenkets.[50] Hence, this symmetrization procedure provides us with all possible gauge-invariant tight-binding models.

The orbitals in Eq. (2.10) are useful as a starting point, but they cannot be interpreted as atomic orbitals because they do not have inversion symmetry. To obtain more "atomic-like" orbitals, one can apply the symmetry operations of the cubic group O_h to the basis ket $|ccc\rangle$, which yields the orbitals

$$\begin{aligned}
|\Gamma_1\rangle &= |s\rangle\\
&= \frac{1}{\sqrt{8}}(|ccc\rangle + |c\bar{c}\bar{c}\rangle + |\bar{c}c\bar{c}\rangle + |\bar{c}\bar{c}c\rangle\\
&\qquad +|\bar{c}\bar{c}\bar{c}\rangle + |\bar{c}cc\rangle + |c\bar{c}c\rangle + |cc\bar{c}\rangle),\\
|\Gamma_{2'}\rangle &= |f_{xyz}\rangle\\
&= \frac{1}{\sqrt{8}}(|ccc\rangle + |c\bar{c}\bar{c}\rangle + |\bar{c}c\bar{c}\rangle + |\bar{c}\bar{c}c\rangle\\
&\qquad -|\bar{c}\bar{c}\bar{c}\rangle - |\bar{c}cc\rangle - |c\bar{c}c\rangle - |cc\bar{c}\rangle),\\
|\Gamma_{15}^z\rangle &= |p_z\rangle\\
&= \frac{1}{\sqrt{8}}(|ccc\rangle - |c\bar{c}\bar{c}\rangle - |\bar{c}c\bar{c}\rangle + |\bar{c}\bar{c}c\rangle\\
&\qquad -|\bar{c}\bar{c}\bar{c}\rangle + |\bar{c}cc\rangle + |c\bar{c}c\rangle - |cc\bar{c}\rangle),\\
|\Gamma_{25'}^{xy}\rangle &= |d_{xy}\rangle\\
&= \frac{1}{\sqrt{8}}(|ccc\rangle - |c\bar{c}\bar{c}\rangle - |\bar{c}c\bar{c}\rangle + |\bar{c}\bar{c}c\rangle\\
&\qquad +|\bar{c}\bar{c}\bar{c}\rangle - |\bar{c}cc\rangle - |c\bar{c}c\rangle + |cc\bar{c}\rangle).
\end{aligned} \tag{2.11}$$

Here two different labels are used: the representations of O_h,[55] and the conventional atomic orbital notation. Other orbitals not given here may be obtained from cyclic permutations of x, y, and z. Note that the T_d orbital $|\Gamma_{15}^z\rangle$ in Eq. (2.10) is the same as $\frac{1}{\sqrt{2}}(|p_z\rangle + |d_{xy}\rangle)$, whereas the T_d orbital $|\Gamma_1\rangle$ is just $\frac{1}{\sqrt{2}}(|s\rangle + |f_{xyz}\rangle)$.

In the basis (2.11), the nonzero coordinate matrix elements are

$$\langle s|x|p_x\rangle = \langle p_x|y|d_{xy}\rangle = \langle d_{xy}|z|f_{xyz}\rangle = c, \tag{2.12}$$

plus others given by cyclic permutations. The selection rules for $\mathbf{x}$ are thus the same as those in a spherically symmetric atom—although, in any real atom, the matrix elements (2.12) would not be numerically equal. This equality occurs because the basis kets (2.11) are degenerate eigenkets of the radial coordinate $r = \sqrt{x^2+y^2+z^2}$. To break the numerical equality, one would need to use basis functions with a linear combination of different radii.

The above procedure may, of course, be applied to coordinate eigenkets other than $|ccc\rangle$. Below is a list of the representations obtained by applying the symmetry operations of O_h to several different "generator" eigenkets:

$$\begin{aligned}
|000\rangle &\rightarrow \Gamma_1\\
&\rightarrow s,\\
|100\rangle &\rightarrow \Gamma_1 + \Gamma_{15} + \Gamma_{12}\\
&\rightarrow s^1p^3d^2,\\
|111\rangle &\rightarrow \Gamma_1 + \Gamma_{15} + \Gamma_{25'} + \Gamma_{2'}\\
&\rightarrow s^1p^3d^3f^1,\\
|110\rangle &\rightarrow \Gamma_1 + \Gamma_{12} + \Gamma_{15} + \Gamma_{25'} + \Gamma_{25}\\
&\rightarrow s^1p^3d^5f^3.
\end{aligned} \tag{2.13}$$

Explicit basis functions for these representations are given in Appendix A.

With these results, one can now construct a gauge-invariant tight-binding model simply by putting a set (or more than one set) of these "orbitals" on each atom in a crystal or molecule. In such a model, the coordinate operators x, y, and z commute by construction. However, one is no longer permitted to choose orbitals arbitrarily. The choices are limited to taking *all* of the orbitals in a set or taking *none* of them. As an example, one cannot discard the f orbitals in the basis generated by $|110\rangle$ without destroying the gauge invariance of the theory.

This approach yields a tight-binding model with orthogonal orbitals. Another approach is to define a grid of coordinates, some points of which are not uniquely associated with individual atoms. One may still construct symmetrized orbitals in this case, but the orbitals are not orthogonal. This makes the tight-binding approach more difficult; however, one can simplify the theory by choosing a Bravais lattice for the coordinate grid, in which case the model may be viewed as a discrete pseudopotential model. Applications of both the pseudopotential and tight-binding approaches are considered below in Sec. V.

III. TOPOLOGY OF THE LATTICE

As we have seen, the most general gauge-invariant tight-binding basis consists of a set of discrete coordinate eigenkets, which will be referred to as a lattice. Such a

lattice is generally not periodic. In order to apply the principle of local gauge symmetry to such a system, one must be able to calculate the change in phase that occurs along any specified path in coordinate space.[36,38] Thus, the first step is to define what is meant by a "path" in a discrete coordinate system.

In general, a path is just an ordered sequence of points. In a continuous coordinate system, neighboring points in the sequence must be separated by an infinitesimal distance. This defines the *topology* of the system, in which points are linked together only if they are adjacent in coordinate space. It is desirable to define the topology of the discrete lattice in a similar way.

One way to do this is to construct a Voronoi polyhedron around each site in the lattice.[47] A Voronoi polyhedron is just the region in space closest to that point;[56] if the lattice is a Bravais lattice, the polyhedron is the same as a Wigner-Seitz cell. In mathematical terms, the Voronoi polyhedron Ω_i for site $\mathbf{x}_i$ is the set of points $\mathbf{x}$ such that $|\mathbf{x}-\mathbf{x}_i| \leq |\mathbf{x}-\mathbf{x}_j|$ for all $j \neq i$. The topology is then defined by the following rule: If Ω_i and Ω_j share a surface with area $S_{ij} > 0$, the sites $\mathbf{x}_i$ and $\mathbf{x}_j$ are linked together; otherwise, they are not.

In some cases, it may happen that two Voronoi polyhedra share only a point or a line, in which case $S_{ij} = 0$. The linking algorithm presented in Ref. 47 does not consider this possibility (because Ref. 47 deals only with random lattices, for which the probability of such an event is zero). For certain applications, it is useful to include links between such sites,[56] but we shall see below that these links should be excluded in the present situation. Thus, only adjacent sites whose Voronoi polyhedra share a surface with $S_{ij} > 0$ are linked.

A path in the discrete lattice is then just an ordered sequence of linked points, and a closed path is one whose first and last points are the same. By definition, every edge of a Voronoi polyhedron is equidistant from three or more lattice sites, all of which lie in a plane perpendicular to the given edge. These sites are closer to this edge than any other sites. The links between these sites form a closed path, and the area bordered by the links is called a "plaquette." There is a one-to-one relationship between the plaquettes and the edges of the Voronoi polyhedra.

The plaquettes partition all of coordinate space into nonoverlapping volumes called (Delaunay) cells. Each cell is uniquely associated with one corner of a Voronoi polyhedron. The partitioning of space into cells is referred to as a Delaunay tessellation.[56]

A general algorithm for calculating the geometry of Voronoi polyhedra, links, plaquettes, and cells is presented in Appendix B. The expressions derived there will be of use in what follows.

IV. LOCAL GAUGE SYMMETRY ON AN ARBITRARY DISCRETE LATTICE

Christ, Friedberg, and Lee[47,48,49] have developed a theory of local gauge symmetry on a random lattice. This section presents a modified version of their theory, with special emphasis on the implications of the principle of local gauge symmetry for tight-binding theory. The presentation follows Dirac's approach,[36] in which the existence of electromagnetic fields is "derived" as a straightforward consequence of a degree of freedom (nonintegrable phases) possessed by any quantum-mechanical system that can be represented in a coordinate basis.

A. Electromagnetism is a nonintegrable phase

In a discrete coordinate basis, any ket vector may be expressed as

$$|\psi\rangle = \sum_i c_i |\mathbf{x}_i\rangle, \tag{4.1}$$

where $|\mathbf{x}_i\rangle$ is a coordinate eigenket, which is assumed to be normalized such that

$$\langle \mathbf{x}_i | \mathbf{x}_j \rangle = \delta_{ij}. \tag{4.2}$$

Dirac's starting point[36] is the fact that physical predictions in quantum mechanics are ultimately expressed in terms of probabilities of the form $|\langle\psi|\psi'\rangle|^2$, where the probability amplitude $\langle\psi|\psi'\rangle$ is given by

$$\langle\psi|\psi'\rangle = \sum_i c_i^* c_i'. \tag{4.3}$$

The probability is obviously well defined even when the overall phase of $|\psi\rangle$ has no definite value. This degree of freedom is referred to as global gauge symmetry.

The existence of global gauge symmetry raises the question of whether it is necessary for the local probability amplitude c_i to have a definite phase. In other words, suppose we write

$$c_i = b_i e^{i\beta_i}, \tag{4.4}$$

where the phase of b_i is well defined (to within an integer multiple of 2π), but β_i is a nonintegrable function—that is, the change in β_i around a closed path can take on any value. In this case one can see that $|\langle\psi|\psi'\rangle|^2$ is well defined *only* if the change in β_i around any closed path is the *same* for all states $|\psi\rangle$ and $|\psi'\rangle$ (to within an integer multiple of 2π, which is absorbed into the definition of b_i). But anything that is the same for all states can be viewed as a physically real part of the dynamical system. Since the present system consists only of a single point particle, these nonintegrable phases must represent a field of force acting on the particle.

The principle of local gauge symmetry is therefore defined by the following two postulates:[36] (i) The physical predictions of the theory must be unambiguous. (ii) The phase of c_i at any point in space and time need not be well defined; only the *change* in phase between *linked*

points must be definite. As shown above, these postulates entail that the change in β_i around any closed path must be the same for all states. According to postulate (ii), this change is fixed for any path by the change in β_i between two linked points in space:

$$\kappa_{ij} = \beta_i - \beta_j \qquad (i \text{ linked to } j), \tag{4.5}$$

and in time:

$$\lambda_i = \frac{\mathrm{d}\beta_i}{\mathrm{d}t} \equiv \dot{\beta}_i. \tag{4.6}$$

Since β_i is nonintegrable, κ_{ij} and λ_i are *independent* variables. These two quantities are the fundamental dynamical variables that arise from the principle of local gauge symmetry. It will now be shown that κ_{ij} and λ_i can be interpreted as *potentials* for the electromagnetic field.

One possible closed path involves a space displacement $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ and an infinitesimal time displacement $\mathrm{d}t$, followed by $\mathbf{d}_{ji}$ and $-\mathrm{d}t$. The change in phase around this path can be used to define (tentatively) an electric field variable

$$E_{ji} = -\frac{\hbar}{e}\frac{\lambda_j - \lambda_i - \dot{\kappa}_{ji}}{d_{ji}}, \tag{4.7a}$$

where $d_{ji} = |\mathbf{d}_{ji}|$. If the index ℓ is used to label the links, one may write this in the simpler form

$$E_\ell = -\frac{\hbar}{e}\frac{\Delta\lambda_\ell - \dot{\kappa}_\ell}{d_\ell}. \tag{4.7b}$$

Here E_ℓ is interpreted as the component of the electric field in the direction $\mathbf{d}_\ell = \mathbf{d}_{ji}$; the components perpendicular to $\mathbf{d}_\ell$ are not defined. Equation (4.7) takes a familiar form if expressed in terms of the potentials

$$\phi_i = \frac{\hbar}{e}\lambda_i, \qquad A_\ell = -\frac{\hbar c}{e}\frac{\kappa_\ell}{d_\ell}, \tag{4.8}$$

since then

$$E_\ell = -\frac{\Delta\phi_\ell}{d_\ell} - \frac{1}{c}\frac{\partial A_\ell}{\partial t}. \tag{4.9}$$

Here the notation $\partial A_\ell/\partial t$ indicates that d_ℓ is to be held constant during the differentiation.

Another type of closed path is an elementary plaquette q constructed from links ℓ in coordinate space (see Sec. III and Appendix B). The change in phase around the perimeter of the plaquette may be used to define the magnetic field

$$B_q = -\frac{1}{S_q}\frac{\hbar c}{e}\sum_{\ell\in q}\kappa_\ell = \frac{1}{S_q}\sum_{\ell\in q}A_\ell d_\ell, \tag{4.10}$$

where S_q is the area of the plaquette [see Eq. (B24)]. Summing Eq. (4.10) over the (closed) surface of a cell c leads immediately to the "no monopoles" law:[57]

$$\sum_{q\in c} B_q S_q = 0. \tag{4.11}$$

Likewise, summing $E_\ell d_\ell$ around the perimeter of a plaquette gives Faraday's law:

$$\sum_{d_\ell\in S_q} E_\ell d_\ell = -\frac{1}{c}\frac{\mathrm{d}}{\mathrm{d}t}(B_q S_q). \tag{4.12}$$

The other two Maxwell equations can be obtained from the Lagrangian $L = L_f + L_e$, where L_f is the electromagnetic field Lagrangian

$$L_f = \frac{1}{8\pi}\sum_\ell 3E_\ell^2\Omega_\ell - \frac{1}{8\pi}\sum_q 3B_q^2\Omega_q. \tag{4.13}$$

Here $\Omega_\ell = \frac{1}{3}S_\ell d_\ell$ is the volume of link ℓ, where $S_\ell = S_{ij}$ is the area of the surface shared by the Voronoi polyhedra for sites i and j (see Appendix B). Likewise, $\Omega_q = \frac{1}{3}S_q d_q$ is the volume of plaquette q, where d_q is the length of the Voronoi polyhedron edge corresponding to q. Equation (4.13) is just a discrete version of the standard field Lagrangian[58] $\frac{1}{8\pi}\int(E^2 - B^2)\mathrm{d}^3x$; the only apparent difference is an extra factor of 3. This factor cancels the factor of $\frac{1}{3}$ in the definition of Ω_ℓ and Ω_q, thus leading to the correct Maxwell equations below. It appears in Eq. (4.13) because the standard Lagrangian is expressed in terms of $E^2 = E_x^2 + E_y^2 + E_z^2$, whereas E_ℓ^2 includes only the component of $\mathbf{E}$ in the direction of $\mathbf{d}_\ell$.

The electronic term in the Lagrangian, which includes the field-particle coupling, is

$$L_e = \frac{i\hbar}{2}\sum_i(c_i^*\dot{c}_i - \dot{c}_i^* c_i) - \sum_{i,j} c_i^* H_{ij} c_j \tag{4.14a}$$

$$= \frac{i\hbar}{2}\sum_i(b_i^*\dot{b}_i - \dot{b}_i^* b_i) - \sum_{i,j} b_i^*\tilde{H}_{ij} b_j. \tag{4.14b}$$

Here $H_{ij} = \langle\mathbf{x}_i|H|\mathbf{x}_j\rangle$ is the Hamiltonian in the absence of electromagnetic fields, while

$$\tilde{H}_{ij} = H_{ij}e^{-i(\beta_i - \beta_j)} + \hbar\dot{\beta}_i\delta_{ij}. \tag{4.15}$$

The first expression (4.14a) for L_e has exactly the same form as the Lagrangian in the case of no electromagnetic fields. This expresses the fundamental physical content of the principle of local gauge symmetry—that the influence of the field upon the particle can be expressed *entirely* in terms of the nonintegrable phase of the probability amplitude $c_i = b_i e^{i\beta_i}$. In the second expression (4.14b) for L_e, all of the nonintegrable phases are collected together in the effective Hamiltonian $\tilde{H}_{ij}$. This is the usual approach, in which the probability amplitude b_i has a well-defined phase, and the field appears only in the Hamiltonian.

The Hamiltonian (4.15) appearing in the Lagrangian (4.14b) depends upon the phase difference $\beta_i - \beta_j$. This phase difference is not well defined unless the sites i and j are linked. But according to postulate (i) above, all physical predictions of the theory must be unambiguous. Hence, the principle of local gauge symmetry demands that

$$H_{ij} = 0 \qquad (i \text{ not linked to } j). \tag{4.16}$$

Local gauge symmetry therefore imposes constraints not found in conventional tight-binding models.

Note that the Lagrangian L is gauge invariant by construction. In other words, both L_f and L_e are invariant under the gauge (phase) transformation

$$\begin{aligned} b_i &\rightarrow b_i e^{-i\chi_i}, \\ \lambda_i &\rightarrow \lambda_i + \dot{\chi}_i, \\ \kappa_{ij} &\rightarrow \kappa_{ij} + \chi_i - \chi_j, \end{aligned} \tag{4.17}$$

where χ_i is an arbitrary integrable function.

Given the above Lagrangian, the Euler-Lagrange equation for λ_i or ϕ_i is just Gauss's law

$$\sum_j E_{ji} S_{ji} = 4\pi q_i, \tag{4.18}$$

where $q_i = e b_i^* b_i$ is the charge on site i. The corresponding equation for κ_ℓ is the Ampère-Maxwell equation

$$\sum_{d_q \in S_\ell} B_q d_q - \frac{1}{c}\frac{\mathrm{d}}{\mathrm{d}t}(E_\ell S_\ell) = \frac{4\pi}{c} I_\ell, \tag{4.19}$$

where $I_\ell = I_{ji} = (2e/\hbar)\mathrm{Im}(b_j^* \tilde{H}_{ji} b_i)$ is the current from site i to site j. Summing (4.19) over all links that contain a given site i yields the charge conservation law[59]

$$\dot{q}_i + \sum_j I_{ji} = 0. \tag{4.20}$$

Thus, we see that λ_i and κ_ℓ can be given a consistent interpretation as discrete electromagnetic potentials, since the above equations are in full agreement with macroscopic (i.e., long-wavelength) electromagnetism.

In some applications of Voronoi polyhedra, it is useful to link sites i and j whose polyhedra share only a line or point, hence $S_\ell = S_{ij} = 0$.[56] For such links, the link volume $\Omega_\ell = \frac{1}{3} S_\ell d_\ell$ is zero, so the electric field E_ℓ does not contribute to the field Lagrangian (4.13), Gauss's law (4.18), or the Ampère-Maxwell equation (4.19). The magnetic-field contribution to (4.19) likewise vanishes, because $S_\ell = 0$. The current through such a link must therefore be zero, which can only be true in general if the Hamiltonian matrix element H_{ij} vanishes. Links with $S_{ij} = 0$ are consequently devoid of any physical significance, and there is no loss of generality if one excludes them at the outset by linking only sites with $S_{ij} > 0$.

Returning to the Lagrangian L, the Euler-Lagrange equation for b_i^* is just the Schrödinger equation

$$i\hbar \dot{b}_i = \sum_j \tilde{H}_{ij} b_j. \tag{4.21}$$

Since b_i is an ordinary probability amplitude with a well-defined phase, $\tilde{H}_{ij}$ must be the Hamiltonian in the presence of electromagnetic fields. With the restriction (4.16), one can express (4.15) as

$$\begin{aligned} \tilde{H}_{ij} &= H_{ij} \exp(-i\kappa_{ij}) + \hbar\lambda_i \delta_{ij} \\ &= H_{ij} \exp(ieA_{ij} d_{ij}/\hbar c) + e\phi_i \delta_{ij}. \end{aligned} \tag{4.22}$$

Note the strong similarity between this result and the Peierls substitution (1.4). The main difference is that (4.22) gives the Hamiltonian in the coordinate representation, not the tight-binding representation.

If $|\kappa_{ij}| \ll 1$ (i.e., if the field is weak or the lattice spacing is small), the Hamiltonian (4.22) reduces to

$$\tilde{H}_{ij} \simeq H_{ij} - \frac{e}{mc}\mathbf{A}_{ij} \cdot \mathbf{p}_{ij} + \frac{e^2 A_{ij}^2}{2mc^2}\Delta_{ij} + e\phi_i \delta_{ij}. \tag{4.23}$$

Here a vector potential has been defined as $\mathbf{A}_{ij} = A_{ij}\hat{\mathbf{d}}_{ij}$, while the momentum operator is given by

$$\mathbf{p}_{ij} = \frac{m}{i\hbar}\mathbf{d}_{ij} H_{ij} = \frac{m}{i\hbar}(\mathbf{x}_i - \mathbf{x}_j) H_{ij}, \tag{4.24}$$

which is the same as the kinematic momentum defined above in Eq. (1.1). The Hamiltonian (4.22) therefore clearly gives the correct first-order $\mathbf{A} \cdot \mathbf{p}$ coupling. We shall see below that the dimensionless quantity

$$\Delta_{ij} = \frac{1}{i\hbar}\mathbf{d}_{ij} \cdot \mathbf{p}_{ij} = -\frac{m}{\hbar^2} d_{ij}^2 H_{ij} \tag{4.25}$$

can be viewed as a geometric weight factor that gives the correct A^2 coupling also.

B. Geometric definition of momentum and kinetic energy

Up to this point, little has been said about the structure of the Hamiltonian H_{ij}. Within the bounds of the restriction (4.16), H_{ij} may be treated as an arbitrary fitting parameter. However, in some circumstances it may be desirable to reduce the number of fitting parameters by using a theoretical formula for H_{ij} that would reproduce the Schrödinger equation in the limit of zero lattice spacing.

Let us start by considering the momentum operator, which will be defined in this section as the canonical momentum $\mathbf{p} = -i\hbar\nabla$. A discrete expression for the gradient operator may be obtained from the integral definition of the gradient:[60]

$$\nabla f(\mathbf{x}) = \lim_{\Omega \rightarrow 0}\left[\frac{1}{\Omega}\int_{\partial\Omega} f(\mathbf{x})\mathrm{d}\mathbf{S}\right], \tag{4.26}$$

where $\mathrm{d}\mathbf{S}$ is a surface element pointing outward from Ω. Now the limiting volume in a discrete lattice is the volume Ω_i of the Voronoi polyhedron for site $\mathbf{x}_i$. On the surface S_{ij} shared by Ω_i and Ω_j, the value of $f(\mathbf{x})$ may be taken to be $\frac{1}{2}[f(\mathbf{x}_i) + f(\mathbf{x}_j)]$. Hence, the discrete gradient may be defined as

$$\nabla f(\mathbf{x}_i) = \frac{1}{\Omega_i}\sum_j \frac{1}{2}[f(\mathbf{x}_i) + f(\mathbf{x}_j)]\mathbf{S}_{ji}, \tag{4.27}$$

where $\mathbf{S}_{ji} = S_{ji}\hat{\mathbf{d}}_{ji}$. Now since

$$\sum_j \mathbf{S}_{ji} = \int_{\partial\Omega_i} \mathrm{d}\mathbf{S} = 0, \tag{4.28}$$

the term involving $f(\mathbf{x}_i)$ drops out, leaving only

$$\nabla f(\mathbf{x}_i) = \frac{1}{2\Omega_i} \sum_j f(\mathbf{x}_j) \mathbf{S}_{ji}. \tag{4.29}$$

An alternative derivation of this result is given in Eq. (17) of Ref. 49.

The canonical momentum operator $\mathbf{p}$ may therefore be defined as[61]

$$\begin{aligned} \langle \mathbf{X}_i | \mathbf{p} | \varphi \rangle &= -i\hbar \nabla \langle \mathbf{X}_i | \varphi \rangle & \text{(4.30a)} \\ &= -\frac{i\hbar}{2\Omega_i} \sum_j \langle \mathbf{X}_j | \varphi \rangle \mathbf{S}_{ji}. & \text{(4.30b)} \end{aligned}$$

Here the basis kets $|\mathbf{X}_i\rangle = \Omega_i^{-1/2} |\mathbf{x}_i\rangle$ are chosen to satisfy "δ-function" normalization

$$\langle \mathbf{X}_i | \mathbf{X}_j \rangle = \frac{\delta_{ij}}{\Omega_i}, \tag{4.31}$$

in contrast to the usual kets $|\mathbf{x}_i\rangle$, which are normalized to unity [see Eq. (4.2)]. The normalization (4.31) is used here because it agrees (in the limit $\Omega_i \to 0$) with the δ-function normalization of continuous coordinate eigenkets, upon which the definition (4.30a) is based.[61]

Substituting $|\varphi\rangle = |\mathbf{X}_j\rangle$ in Eq. (4.30b) then gives

$$\langle \mathbf{X}_i | \mathbf{p} | \mathbf{X}_j \rangle = \frac{i\hbar \mathbf{S}_{ij}}{2\Omega_i \Omega_j}, \tag{4.32}$$

which clearly satisfies

$$\sum_j \langle \mathbf{X}_i | \mathbf{p} | \mathbf{X}_j \rangle \Omega_j = 0. \tag{4.33}$$

Replacing $|\mathbf{X}_i\rangle = \Omega_i^{-1/2} |\mathbf{x}_i\rangle$ in (4.32) then yields the desired result

$$\langle \mathbf{x}_i | \mathbf{p} | \mathbf{x}_j \rangle = \frac{i\hbar \mathbf{S}_{ij}}{2\sqrt{\Omega_i \Omega_j}}. \tag{4.34}$$

Note that this matrix is Hermitian, because $\mathbf{S}_{ji} = -\mathbf{S}_{ij}$. If the kets $|\mathbf{x}_i\rangle$ are used in Eq. (4.30a) above, a non-Hermitian canonical momentum is obtained.

There is some question as to whether this definition of $\mathbf{p}$ should be referred to as "canonical," because it does not satisfy the canonical commutation relations. In a continuous coordinate basis, the canonical momentum satisfies

$$\langle \mathbf{x}' | [x^\alpha, p^\beta] | \mathbf{x}'' \rangle = i\hbar \delta_{\alpha\beta} \delta(\mathbf{x}' - \mathbf{x}''), \tag{4.35}$$

where α and β are Cartesian components of the given vectors. The corresponding equation in the discrete basis is

$$\langle \mathbf{X}_i | [x^\alpha, p^\beta] | \mathbf{X}_j \rangle = \frac{i\hbar d_{ij}^\alpha S_{ij}^\beta}{2\Omega_i \Omega_j}, \tag{4.36}$$

which obviously does not agree. Note, however, that

$$\begin{aligned} \sum_{i,j} \langle \mathbf{X}_i | [x^\alpha, p^\beta] | \mathbf{X}_j \rangle \Omega_i \Omega_j &= \frac{i\hbar}{2} \sum_{i,j} d_{ij}^\alpha S_{ij}^\beta \\ &= i\hbar \delta_{\alpha\beta} \Omega, & \text{(4.37)} \end{aligned}$$

where Ω is the total volume, and the second equality is proved in Eq. (11) of Ref. 49. This agrees with the relation

$$\iint \langle \mathbf{x}' | [x^\alpha, p^\beta] | \mathbf{x}'' \rangle \mathrm{d}^3x' \mathrm{d}^3x'' = i\hbar \delta_{\alpha\beta} \Omega \tag{4.38}$$

in the continuous basis. Hence, Eq. (4.37) is as close as one can come to a canonical commutation relation in a general discrete basis.[62,63]

A similar definition may be used for the kinetic energy operator $T = -\hbar^2 \nabla^2 / 2m$. The integral definition of the divergence,[60]

$$\nabla \cdot \mathbf{F}(\mathbf{x}) = \lim_{\Omega \to 0} \left[\frac{1}{\Omega} \int_{\partial\Omega} \mathbf{F}(\mathbf{x}) \cdot \mathrm{d}\mathbf{S} \right], \tag{4.39}$$

gives the Laplacian

$$\nabla^2 f(\mathbf{x}) = \lim_{\Omega \to 0} \left[\frac{1}{\Omega} \int_{\partial\Omega} \nabla f(\mathbf{x}) \cdot \mathrm{d}\mathbf{S} \right], \tag{4.40}$$

the discrete form of which is

$$\nabla^2 f(\mathbf{x}_i) = \frac{1}{\Omega_i} \sum_j \left(\frac{f(\mathbf{x}_j) - f(\mathbf{x}_i)}{d_{ji}} \right) S_{ji}. \tag{4.41}$$

An alternative derivation of this result is given in Eq. (12) of Ref. 49. The procedure used above in Eqs. (4.30)–(4.34) then yields the kinetic energy operator

$$\langle \mathbf{x}_i | T | \mathbf{x}_j \rangle = -\frac{\hbar^2}{2m} \frac{S_{ij}}{d_{ij} \sqrt{\Omega_i \Omega_j}} + \delta_{ij} \frac{\hbar^2}{2m\Omega_i} \sum_k \frac{S_{ik}}{d_{ik}}, \tag{4.42}$$

which satisfies [cf. Eq. (4.33)]

$$\sum_j \langle \mathbf{X}_i | T | \mathbf{X}_j \rangle \Omega_j = 0. \tag{4.43}$$

Note that for $i \neq j$, $\langle \mathbf{x}_i | T | \mathbf{x}_j \rangle$ decreases continuously to zero when $S_{ij} \to 0$. This ensures that the Hamiltonian is a continuous function of the lattice coordinates, even as new links are formed and old ones are broken.

Such continuity is also desirable when the Hamiltonian is determined empirically, especially for applications (such as molecular dynamics[5,64]) in which the atomic positions vary with time. This can be achieved by defining the nonlocal elements of the empirical Hamiltonian as

$$\langle \mathbf{x}_i | H | \mathbf{x}_j \rangle = -\frac{\hbar^2}{2m} \frac{S_{ij}}{d_{ij} \sqrt{\Omega_i \Omega_j}} f_{ij} \qquad (i \neq j), \tag{4.44}$$

where the fitting parameter f_{ij} is a continuous, nonsingular function of the lattice coordinates.

Note that the operators $\mathbf{p}$ and T do not satisfy $T = \mathbf{p}^2/2m$, because $\mathbf{p}^2$, unlike $\mathbf{p}$ and T, couples sites that are not linked. However, $\mathbf{p}$ and T are related by

$$\mathbf{p}_{ij} = \frac{m}{i\hbar}\mathbf{d}_{ij}T_{ij} \tag{4.45a}$$

or

$$\mathbf{p} = \frac{m}{i\hbar}[\mathbf{x}, T]. \tag{4.45b}$$

Thus, any Hamiltonian of the form $H = T + V(\mathbf{x})$, where V is a local potential, satisfies Eq. (1.1). Hence, for such a Hamiltonian, the canonical momentum $\mathbf{p} = -i\hbar\nabla$ used in this section agrees with the kinematic momentum defined earlier.

Now let us examine the dimensionless factor Δ_{ij} defined above in Eq. (4.25). If $H = T + V$, this becomes

$$\Delta_{ij} = \frac{S_{ij}d_{ij}}{2\sqrt{\Omega_i\Omega_j}} = \frac{3\Omega_{ij}}{2\sqrt{\Omega_i\Omega_j}}, \tag{4.46}$$

where $\Omega_{ij} = \frac{1}{3}S_{ij}d_{ij}$ is the volume of the link between sites i and j (see Appendix B). The factor Δ_{ij} appears in the A^2 term in the Hamiltonian (4.23), which will be referred to as H_2. In a continuous coordinate basis, H_2 is given by

$$\langle \mathbf{x}'|H_2|\mathbf{x}''\rangle = \frac{e^2}{2mc^2}A^2(\mathbf{x}')\delta(\mathbf{x}' - \mathbf{x}''), \tag{4.47}$$

which means that it satisfies

$$\iint \langle \mathbf{x}'|H_2|\mathbf{x}''\rangle \mathrm{d}^3x'\mathrm{d}^3x'' = \frac{e^2}{2mc^2}\int A^2(\mathbf{x}')\mathrm{d}^3x'. \tag{4.48}$$

The corresponding equations for the discrete basis are

$$\langle \mathbf{X}_i|H_2|\mathbf{X}_j\rangle = \frac{e^2}{2mc^2}\frac{3A_{ij}^2\Omega_{ij}}{2\Omega_i\Omega_j} \tag{4.49}$$

and

$$\begin{aligned}\sum_{i,j}\langle \mathbf{X}_i|H_2|\mathbf{X}_j\rangle\Omega_i\Omega_j &= \frac{e^2}{2mc^2}\frac{3}{2}\sum_{i,j}A_{ij}^2\Omega_{ij} \\ &= \frac{e^2}{2mc^2}\sum_{\ell}3A_\ell^2\Omega_\ell. \end{aligned}\tag{4.50}$$

The second equality in (4.50) was obtained by noting that a sum over i and j covers each link ℓ twice. The only apparent difference between Eqs. (4.48) and (4.50) is a factor of 3. This appears for the same reason that it does in the Lagrangian (4.13)—i.e., A^2 in Eq. (4.48) refers to $A_x^2 + A_y^2 + A_z^2$, whereas A_ℓ^2 in Eq. (4.50) refers only to the component of $\mathbf{A}$ in the direction of $\mathbf{d}_\ell$.

Therefore, Eqs. (4.48) and (4.50) are the same in the limit of zero lattice spacing, and the factor Δ_{ij} is simply a geometric weight factor that provides the correct A^2 coupling in the Hamiltonian (4.23).

C. Spin

The theory presented thus far has been for a particle with spin zero. Particles with spin $\frac{1}{2}$ may be described using a discrete version of the Dirac Hamiltonian for a free particle:

$$H = c\boldsymbol{\alpha}\cdot\mathbf{p} + \beta mc^2, \tag{4.51}$$

where $\boldsymbol{\alpha}$ and β are Dirac's 4×4 matrices. The momentum operator $\mathbf{p}$ can either be calculated from geometry or fitted to experiment. In the presence of electromagnetic fields, the Hamiltonian becomes

$$\tilde{H} = c\boldsymbol{\alpha}\cdot\boldsymbol{\pi} + \beta mc^2 + e\phi, \tag{4.52}$$

where [cf. Eq. (4.22)]

$$\boldsymbol{\pi}_{ij} = \mathbf{p}_{ij}\exp(-i\kappa_{ij}). \tag{4.53}$$

A nonrelativistic Hamiltonian may be obtained by applying a Foldy-Wouthuysen transformation[65,66] to Eq. (4.52), which yields

$$\begin{aligned}H_{\mathrm{nr}} &= \frac{1}{2m}(\boldsymbol{\sigma}\cdot\boldsymbol{\pi})^2 - \frac{1}{8m^3c^2}(\boldsymbol{\sigma}\cdot\boldsymbol{\pi})^4 + e\phi \\ &\quad - \frac{1}{8m^2c^2}[\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, ([\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, e\phi] + i\hbar\boldsymbol{\sigma}\cdot\dot{\boldsymbol{\pi}})],\end{aligned}\tag{4.54}$$

where $\boldsymbol{\sigma}$ is the Pauli spin matrix, and all terms of order $(v/c)^4$ have been included. This Hamiltonian couples sites that are not linked, but there is no ambiguity because the Dirac equation is taken as fundamental.

If we assume for simplicity that the lattice coordinates do not depend on time, then

$$([\boldsymbol{\sigma}\cdot\boldsymbol{\pi}, e\phi] + i\hbar\boldsymbol{\sigma}\cdot\dot{\boldsymbol{\pi}})_{ij} = -\boldsymbol{\sigma}\cdot\boldsymbol{\pi}_{ij}V_{ij}, \tag{4.55}$$

where

$$V_{ij} = e(\phi_i - \phi_j) - \hbar\dot{\kappa}_{ij} = -eE_{ij}d_{ij} \tag{4.56}$$

is the difference in potential energy of sites i and j due to the electric field. The last term in Eq. (4.54) therefore consists of the Darwin term

$$H_{ij}^{\mathrm{D}} = \frac{1}{8m^2c^2}\sum_k(V_{ki} + V_{kj})\boldsymbol{\pi}_{ik}\cdot\boldsymbol{\pi}_{kj} \tag{4.57}$$

plus the spin-orbit coupling

$$H_{ij}^{\mathrm{so}} = \frac{i}{8m^2c^2}\sum_k(V_{ki} + V_{kj})\boldsymbol{\sigma}\cdot(\boldsymbol{\pi}_{ik}\times\boldsymbol{\pi}_{kj}), \tag{4.58}$$

where the identity

$$(\boldsymbol{\sigma}\cdot\mathbf{a})(\boldsymbol{\sigma}\cdot\mathbf{b}) = \mathbf{a}\cdot\mathbf{b} + i\boldsymbol{\sigma}\cdot(\mathbf{a}\times\mathbf{b}) \tag{4.59}$$

has been used. Now the main contribution to spin-orbit coupling comes from the atomic cores, where the potential energy and wave function vary rapidly. However, in

any basis of reasonable size, the lattice imposes a wavelength cutoff that eliminates such rapid variations. The potential ϕ_i must therefore be viewed as a pseudopotential, not a true atomic potential. Hence, for practical purposes, ϕ_i in the spin-orbit Hamiltonian (4.58) should be treated as a fitting parameter that is independent of the value used for the first $e\phi$ term in (4.54).

The first two terms in the Hamiltonian (4.54) are kinetic-energy terms, which may be rewritten using

$$(\boldsymbol{\sigma}\cdot\boldsymbol{\pi})_{ij}^2 = \sum_k [\boldsymbol{\pi}_{ik}\cdot\boldsymbol{\pi}_{kj} + i\boldsymbol{\sigma}\cdot(\boldsymbol{\pi}_{ik}\times\boldsymbol{\pi}_{kj})], \qquad (4.60)$$

in which the second term describes the intrinsic magnetic dipole moment of the particle. For a general lattice, this term is not zero even when there is no electromagnetic field, because different components of the momentum operator do not commute (i.e., $\mathbf{p}\times\mathbf{p}\neq 0$). This follows from the fact that there is generally no more than one site k linked to both i and j, and for that i and j, $\mathbf{p}_{ik}\times\mathbf{p}_{kj}$ is generally not zero.

However, if the lattice is a Bravais lattice, then $\mathbf{p}\times\mathbf{p}$ is always zero. This follows from the fact that every site in a Bravais lattice is identical, so for a given nonzero $\mathbf{p}_{ik}\times\mathbf{p}_{kj}$, there is always another site l with $\mathbf{p}_{il}=\mathbf{p}_{kj}$ and $\mathbf{p}_{lj}=\mathbf{p}_{ik}$, hence $\mathbf{p}_{ik}\times\mathbf{p}_{kj}+\mathbf{p}_{il}\times\mathbf{p}_{lj}=0$. Clearly one also has $\mathbf{d}_{ik}\times\mathbf{d}_{kj}+\mathbf{d}_{il}\times\mathbf{d}_{lj}=0$, so the sites i, k, j, and l lie in a single plane. If i is not linked to j and k is not linked to l, then i, k, j, and l form a single plaquette q, which has the shape of a rectangle. Otherwise, they form two triangular plaquettes.

If the momentum operator is given by Eq. (4.34), then the intrinsic magnetic dipole term in the Hamiltonian (4.54) for such a Bravais lattice is

$$H_{ij}^{\mathrm{mag}} = -\frac{e\hbar}{8mc}\left(\frac{S_{ik}S_{kj}S_q^2}{d_{ik}d_{kj}\Omega_i^2}\right)\boldsymbol{\sigma}\cdot\mathbf{B}_q, \qquad (4.61)$$

where the weak-field approximation $|\kappa_{ij}|\ll 1$ has been used, and the direction of $\mathbf{B}_q$ is that of $\mathbf{S}_q$. If the sites i, k, j, and l form a single rectangular plaquette, then S_q is the area of that plaquette; otherwise, it is the combined area of the two triangles (in which case $\mathbf{B}_q$ is the average magnetic field of the two plaquettes).

As an example, consider a simple cubic lattice with lattice constant a, for which $S_{ik}=S_{kj}=S_q=a^2$, $d_{ik}=d_{kj}=a$, and $\Omega_i=a^3$. In this case, the factor in parentheses in Eq. (4.61) is unity, and H_{ij}^{mag} couples sites on opposite corners of each plaquette (with $d_{ij}=\sqrt{2}a$). By comparison, the dipole term in the continuum Hamiltonian is given by

$$\frac{1}{2m}[\boldsymbol{\sigma}\cdot(\mathbf{p}-e\mathbf{A}/c)]^2 = \frac{1}{2m}(\mathbf{p}-e\mathbf{A}/c)^2 - \frac{e\hbar}{2mc}\boldsymbol{\sigma}\cdot\mathbf{B}. \qquad (4.62)$$

The numerical factor in front of this dipole coupling is four times larger than that in Eq. (4.61). This occurs because (4.61) couples each site i to four other sites j.

V. APPLICATION TO TETRAHEDRAL SEMICONDUCTORS

This section considers several different methods of implementing the theory developed in Sec. IV. Spin is neglected in all of the applications that follow.

A. Discrete pseudopotential method

The simplest geometry occurs when the lattice sites $\mathbf{x}_i$ are chosen to lie on a Bravais lattice. One possible approach in this case is to use the geometric expression (4.42) for the kinetic energy T, and assume that the potential energy V is local. This approach will be referred to as the discrete pseudopotential method.

If $\mathbf{x}$ lies on a Bravais lattice (the subscript i is omitted here), one may define a reciprocal lattice as the set of all vectors $\mathbf{g}$ such that $\mathbf{g}\cdot\mathbf{x}=2\pi\times$ integer. The volume of a primitive cell in the direct lattice is denoted ω_0, while that of a primitive cell in the reciprocal lattice is $\omega_0^*=(2\pi)^3/\omega_0$.

In a crystal, the Hamiltonian will be periodic with respect to some larger Bravais lattice whose sites are denoted by $\mathbf{R}$, where $\mathbf{R}\in\{\mathbf{x}\}$. One may then define a corresponding reciprocal lattice as the set of all vectors $\mathbf{G}$ such that $\mathbf{G}\cdot\mathbf{R}=2\pi\times$ integer. The volume of a primitive cell in $\mathbf{R}$ space is $\Omega_0=n\omega_0$, where n is an integer, and the volume of a primitive cell in $\mathbf{G}$ space is $\Omega_0^*=(2\pi)^3/\Omega_0$.

Periodic boundary conditions may then be implemented over an even larger Bravais lattice whose sites are denoted by $\mathbf{L}$, where $\mathbf{L}\in\{\mathbf{R}\}$. The corresponding reciprocal lattice vectors are denoted $\mathbf{k}$, where $\mathbf{k}\cdot\mathbf{L}=2\pi\times$ integer. The volume over which periodic boundary conditions are applied is $\Omega=N\Omega_0$, where N is an integer, and the volume of a primitive cell in $\mathbf{k}$ space is $\Omega^*=(2\pi)^3/\Omega$. Note that according to the above definitions, $\mathbf{G}\in\{\mathbf{k}\}$ and $\mathbf{g}\in\{\mathbf{G}\}$.

Coordinate eigenkets in this system are denoted $|\mathbf{x}\rangle$, where $|\mathbf{x}+\mathbf{L}\rangle\equiv|\mathbf{x}\rangle$ due to the periodic boundary conditions. The orthogonality and closure relations in this basis are therefore

$$\langle\mathbf{x}|\mathbf{x}'\rangle = \sum_{\mathbf{L}}\delta_{\mathbf{x}-\mathbf{x}',\mathbf{L}}, \qquad (5.1)$$

$$\sum_{\mathbf{x}\in\Omega}|\mathbf{x}\rangle\langle\mathbf{x}| = 1. \qquad (5.2)$$

In a system with periodic boundary conditions, the coordinate operator is not well defined; only periodic functions of the coordinate are permitted. Therefore, the definition of the kinematic momentum must be slightly modified. Instead of (1.1), one has

$$\langle\mathbf{x}|\mathbf{p}|\mathbf{x}'\rangle = \frac{m}{i\hbar}(\mathbf{x}-\mathbf{x}')\langle\mathbf{x}|H|\mathbf{x}'\rangle \qquad \text{if } \mathbf{x}\in\Omega \text{ and } \mathbf{x}'\in\Omega, \qquad (5.3)$$

with $\langle \mathbf{x}|\mathbf{p}|\mathbf{x}'\rangle = \langle \mathbf{x}+\mathbf{L}|\mathbf{p}|\mathbf{x}'\rangle$ otherwise.

Another useful representation is the crystal momentum representation

$$|\mathbf{k}\rangle = \frac{1}{\sqrt{\mathcal{N}}}\sum_{\mathbf{x}\in\Omega} e^{i\mathbf{k}\cdot\mathbf{x}}|\mathbf{x}\rangle, \tag{5.4}$$

where $\mathcal{N} = nN = \Omega/\omega_0$. The corresponding orthogonality and closure relations are

$$\langle \mathbf{k}|\mathbf{k}'\rangle = \sum_{\mathbf{g}} \delta_{\mathbf{k}-\mathbf{k}',\mathbf{g}}, \tag{5.5}$$

$$\sum_{\mathbf{k}\in\omega_0^*} |\mathbf{k}\rangle\langle\mathbf{k}| = 1. \tag{5.6}$$

In the crystal momentum representation, a periodic Hamiltonian $\langle \mathbf{x}|H|\mathbf{x}'\rangle = \langle \mathbf{x}+\mathbf{R}|H|\mathbf{x}'+\mathbf{R}\rangle$ couples only those states that differ by a reciprocal lattice vector $\mathbf{G}$:

$$\langle \mathbf{k}'|H|\mathbf{k}\rangle = \sum_{\mathbf{G}} \delta_{\mathbf{k}',\mathbf{k}+\mathbf{G}}\langle \mathbf{k}+\mathbf{G}|H|\mathbf{k}\rangle, \tag{5.7}$$

where

$$\langle \mathbf{k}+\mathbf{G}|H|\mathbf{k}\rangle = \frac{1}{n}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{x}'\in\Omega_0} e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{x}'}\langle \mathbf{x}'|H|\mathbf{x}\rangle e^{i\mathbf{k}\cdot\mathbf{x}}. \tag{5.8}$$

The kinematic momentum (5.3) also satisfies Eq. (5.7). Its matrix elements that are related to those of H by

$$\langle \mathbf{k}+\mathbf{G}|\mathbf{p}|\mathbf{k}\rangle = \frac{m}{\hbar}\nabla_{\mathbf{k}}\langle \mathbf{k}+\mathbf{G}|H|\mathbf{k}\rangle. \tag{5.9}$$

On a Bravais lattice, the kinetic energy (4.42) and canonical momentum (4.34) are translationally invariant:

$$\langle \mathbf{x}|T|\mathbf{x}'\rangle = T(\mathbf{x}-\mathbf{x}'), \tag{5.10}$$

where $T(\mathbf{x}+\mathbf{L}) = T(\mathbf{x})$. The matrix elements of T are therefore given by

$$\langle \mathbf{k}|T|\mathbf{k}'\rangle = T(\mathbf{k})\sum_{\mathbf{g}} \delta_{\mathbf{k}-\mathbf{k}',\mathbf{g}}, \tag{5.11}$$

where

$$T(\mathbf{k}) = \sum_{\mathbf{x}\in\Omega} T(\mathbf{x})e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{5.12}$$

For tetrahedral semiconductors with the diamond or zincblende structure, it is convenient to use a cubic lattice for the grid $\{\mathbf{x}\}$. Expressions for the link distances d_{ij} and surface areas S_{ij} are given in Appendix C for the simple cubic, body-centered cubic, and face-centered cubic lattices. The resulting kinetic-energy operators given by Eqs. (4.42) and (5.12) are

$$\begin{aligned}
T_{\rm sc}(\mathbf{k}) &= \frac{2\hbar^2}{ma^2}[\sin^2(\tfrac{1}{2}k_x a)+\sin^2(\tfrac{1}{2}k_y a)+\sin^2(\tfrac{1}{2}k_z a)],\\
T_{\rm bcc}(\mathbf{k}) &= \frac{\hbar^2}{2ma^2}\{6[1-\cos(\tfrac{1}{2}k_x a)\cos(\tfrac{1}{2}k_y a)\cos(\tfrac{1}{2}k_z a)]\\
&\quad +[\sin^2(\tfrac{1}{2}k_x a)+\sin^2(\tfrac{1}{2}k_y a)+\sin^2(\tfrac{1}{2}k_z a)]\},\\
T_{\rm fcc}(\mathbf{k}) &= \frac{2\hbar^2}{ma^2}[3-\cos(\tfrac{1}{2}k_y a)\cos(\tfrac{1}{2}k_z a)\\
&\quad -\cos(\tfrac{1}{2}k_z a)\cos(\tfrac{1}{2}k_x a)\\
&\quad -\cos(\tfrac{1}{2}k_x a)\cos(\tfrac{1}{2}k_y a)],
\end{aligned} \tag{5.13}$$

all of which reduce to $T(\mathbf{k}) \simeq \hbar^2k^2/2m$ when $ka \ll 1$. Here a is the lattice constant of the grid $\{\mathbf{x}\}$, which is some integer fraction of the lattice constant a_0 of the crystal lattice $\{\mathbf{R}\}$.

A canonical momentum operator $\mathbf{p}(\mathbf{k})$ corresponding to Eq. (4.34) may be defined in a similar manner. This operator is given by

$$\mathbf{p}(\mathbf{k}) = \frac{m}{\hbar}\nabla_{\mathbf{k}}T(\mathbf{k}), \tag{5.14}$$

which follows from Eq. (4.45). Note that this result is just a special case of the kinematic momentum (5.9).

The matrix elements of a local periodic potential $V(\mathbf{x}) = V(\mathbf{x}+\mathbf{R})$ are given by Eq. (5.7), where

$$\langle \mathbf{k}+\mathbf{G}|V|\mathbf{k}\rangle = \frac{1}{n}\sum_{\mathbf{x}\in\Omega_0} V(\mathbf{x})e^{-i\mathbf{G}\cdot\mathbf{x}} \tag{5.15}$$

is independent of $\mathbf{k}$. It will be assumed here that V is a superposition of local atomic pseudopotentials:

$$V(\mathbf{x}) = \sum_{\mu=1}^{N_a}\sum_{\mathbf{R}\in\Omega} v_\mu(\mathbf{x}-\mathbf{R}-\boldsymbol{\tau}_\mu), \tag{5.16}$$

where $v_\mu(\mathbf{x}) = v_\mu(\mathbf{x}+\mathbf{L})$ is the pseudopotential for atom μ, whose position in the unit cell Ω_0 is given by $\boldsymbol{\tau}_\mu$. In this case

$$\langle \mathbf{k}+\mathbf{G}|V|\mathbf{k}\rangle = \frac{1}{N_a}\sum_{\mu=1}^{N_a} v_\mu(\mathbf{G})e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_\mu}, \tag{5.17}$$

where $v_\mu(\mathbf{G})$ is the atomic form factor

$$v_\mu(\mathbf{G}) = \frac{N_a}{n}\sum_{\mathbf{x}\in\Omega} v_\mu(\mathbf{x})e^{-i\mathbf{G}\cdot\mathbf{x}}, \tag{5.18}$$

and N_a is the number of atoms in the unit cell Ω_0.

The main practical difficulty in implementing the discrete pseudopotential method is that $T(\mathbf{k})$ is not a good approximation to the continuum kinetic energy

$$T_{\rm cont}(\mathbf{k}) = \frac{\hbar^2k^2}{2m} \tag{5.19}$$

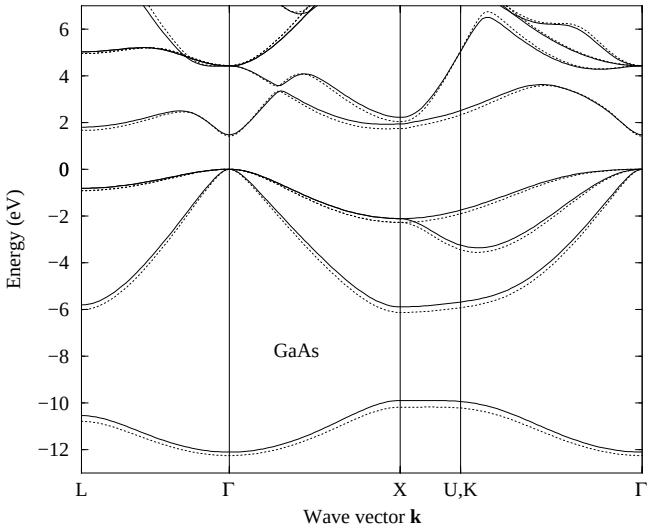

FIG. 1: Energy band structure of GaAs calculated by the discrete local pseudopotential method using an fcc grid with $a = \frac{1}{8}a_0$. Solid lines: fcc kinetic energy from Eq. (5.13). Dotted lines: continuum kinetic energy from Eq. (5.19).

unless $ka \ll 1$. This means that the lattice constant a of the grid $\{\mathbf{x}\}$ must be significantly smaller than the lattice constant a_0 of the crystal lattice $\{\mathbf{R}\}$. To obtain one grid point at each atom in the diamond structure, a must satisfy $a = a_0/2l$ (for a bcc grid) or $a = a_0/4l$ (for sc and fcc), where l is a positive integer. Numerical accuracy generally requires $l > 1$, as shown below.

The *shape* of the Brillouin zone is also important. If the shape of the Wigner-Seitz cell for ω_0^* is not congruent with the shape of the Wigner-Seitz cell for Ω_0^*, then $T(\mathbf{k})$ will deviate from $T_{\rm cont}(\mathbf{k})$ more rapidly in some directions than others. This can lead to significant qualitative errors in the kinetic energy. For example, in diamond, the lowest continuum eigenvalues at Γ are $T_{\rm cont}(\mathbf{G}) = \hbar^2 G^2/2m$, where $\mathbf{G} = (000)$, $\langle 111\rangle$, $\langle 200\rangle$, and $\langle 220\rangle$ (in units of $2\pi/a_0$). But for a sc grid with $a = \frac{1}{4}a_0$, the value of $T_{\rm sc}(\langle 200\rangle) = 2\hbar^2/ma^2$ is actually *lower* than that of $T_{\rm sc}(\langle 111\rangle) = 3\hbar^2/ma^2$. The correct ratio $T_{\rm cont}(\langle 200\rangle) = \frac{4}{3}T_{\rm cont}(\langle 111\rangle)$ is only approached in the limit of very small a/a_0, making the sc grid a poor choice for diamond.

The natural choice for diamond is the fcc grid, since its Brillouin zone has the same shape as that of diamond. Indeed, one has $T_{\rm fcc}(\langle 200\rangle) = \frac{4}{3}T_{\rm fcc}(\langle 111\rangle)$ even for the maximum grid size $a = \frac{1}{4}a_0$. The only problem here is that $T_{\rm fcc}(\langle 220\rangle) = \frac{3}{2}T_{\rm fcc}(\langle 200\rangle)$, which is not sufficiently close to the correct ratio $T_{\rm cont}(\langle 220\rangle) = 2T_{\rm cont}(\langle 200\rangle)$ to make $a = \frac{1}{4}a_0$ a satisfactory choice for numerical calculations. The next possibility is $a = \frac{1}{8}a_0$, which yields $T_{\rm fcc}(\langle 220\rangle) = (\frac{3}{2} + \frac{\sqrt{2}}{4})T_{\rm fcc}(\langle 200\rangle) \simeq 1.85T_{\rm fcc}(\langle 200\rangle)$.

The energy band structure for GaAs calculated using an fcc grid with $a = \frac{1}{8}a_0$ is given in Fig. 1. The fcc kinetic energy obtained from Eq. (5.13) was multiplied by a constant factor $\pi^2(2+\sqrt{2})/32 = 1.053$ so that $T_{\rm fcc}(\mathbf{k})$ matches $T_{\rm cont}(\mathbf{k})$ at $\mathbf{G} = \langle 111\rangle$ and $\mathbf{G} = \langle 200\rangle$. The pseudopotential form factors (5.18) for this calculation were taken from Ref. 67. No attempt was made to fit the energy bands by modifying the form factors; the purpose of this figure is merely to demonstrate the close similarity between the discrete fcc band structure and the continuum band structure. Slight adjustments in the model parameters would likely give an even better agreement.

The main problem with this result is that $a = \frac{1}{8}a_0$ corresponds to a basis size of 512 grid points per primitive unit cell Ω_0. This is unattractive in comparison to the basis dimensions of approximately 100 plane waves that are typically used in empirical pseudopotential calculations for tetrahedral semiconductors. However, changing the fcc grid to $a = \frac{1}{4}a_0$ (i.e., 64 grid points per primitive cell) makes it impossible to achieve a satisfactory fit to the band structure using local pseudopotentials. A good fit is possible only if the nonlocal Hamiltonian matrix elements are treated as fitting parameters. But in that case, one can reduce the basis dimensions even further by using the tight-binding approach.

B. Tight-binding method

In the tight-binding approach, the grid points are no longer restricted to lie on a Bravais lattice, and all of the Hamiltonian matrix elements are treated as fitting parameters. In addition, it is assumed here that the model is constructed using the symmetrization procedure described in Sec. II, so that a distinct set of orthogonal orbitals is associated with each atom. The objective then is to find the smallest coordinate basis that provides a physically reasonable model of a given system.

The basis kets in the tight-binding approach will be written as $|\alpha, \mathbf{R}+\boldsymbol{\tau}_\mu\rangle$, where $\mathbf{R}$ is a lattice vector for the Bravais lattice over which the Hamiltonian is periodic, and $\boldsymbol{\tau}_\mu$ is the position of atom μ within the primitive unit cell Ω_0. These quantities are defined exactly as they were in Sec. V A; the vectors $\mathbf{L}$, $\mathbf{G}$, and $\mathbf{k}$ are also defined in the same way. The label α may refer to a coordinate $\mathbf{x}_\alpha$ within the atom, in which case

$$|\alpha, \mathbf{R}+\boldsymbol{\tau}_\mu\rangle \equiv |\mathbf{x}_\alpha + \mathbf{R} + \boldsymbol{\tau}_\mu\rangle \tag{5.20}$$

is just a coordinate eigenket. However, α may also be used as a symmetry label for an atomic orbital that is a symmetrized linear combination of the kets (5.20):

$$|\alpha, \mathbf{R}+\boldsymbol{\tau}_\mu\rangle = \sum_\beta C_\beta(\alpha)|\mathbf{x}_\beta + \mathbf{R} + \boldsymbol{\tau}_\mu\rangle. \tag{5.21}$$

In either case, the basis is orthogonal:

$$\langle \alpha, \mathbf{R}+\boldsymbol{\tau}_\mu|\alpha', \mathbf{R}'+\boldsymbol{\tau}_{\mu'}\rangle = \delta_{\alpha\alpha'}\delta_{\mu\mu'}\sum_{\mathbf{L}}\delta_{\mathbf{R}-\mathbf{R}',\mathbf{L}}, \tag{5.22}$$

and complete:

$$\sum_{\alpha}\sum_{\mu}\sum_{\mathbf{R}\in\Omega}|\alpha,\mathbf{R}+\boldsymbol{\tau}_\mu\rangle\langle\alpha,\mathbf{R}+\boldsymbol{\tau}_\mu|=1. \tag{5.23}$$

In periodic systems, it is convenient to define the Bloch sums[1]

$$|\alpha,\mu,\mathbf{k}\rangle=\frac{1}{\sqrt{N}}\sum_{\mathbf{R}\in\Omega}e^{i\mathbf{k}\cdot(\mathbf{R}+\boldsymbol{\tau}_\mu)}|\alpha,\mathbf{R}+\boldsymbol{\tau}_\mu\rangle, \tag{5.24}$$

which are also orthogonal and complete:

$$\langle\alpha,\mu,\mathbf{k}|\alpha',\mu',\mathbf{k}'\rangle=\delta_{\alpha\alpha'}\delta_{\mu\mu'}\sum_{\mathbf{G}}\delta_{\mathbf{k}-\mathbf{k}',\mathbf{G}}\,e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_\mu}, \tag{5.25}$$

$$\sum_{\alpha}\sum_{\mu}\sum_{\mathbf{k}\in\Omega_0^*}|\alpha,\mu,\mathbf{k}\rangle\langle\alpha,\mu,\mathbf{k}|=1. \tag{5.26}$$

If the Hamiltonian H is invariant with respect to lattice translations $\mathbf{R}$, then its matrix elements in the Bloch basis are

$$\begin{aligned}\langle\alpha,\mu,\mathbf{k}|H|\alpha',\mu',\mathbf{k}'\rangle &= \langle\alpha,\mu,\mathbf{k}|H|\alpha',\mu',\mathbf{k}\rangle \\ &\quad\times\sum_{\mathbf{G}}\delta_{\mathbf{k}-\mathbf{k}',\mathbf{G}}e^{-i\mathbf{G}\cdot\boldsymbol{\tau}_{\mu'}},\end{aligned} \tag{5.27}$$

where

$$\begin{aligned}\langle\alpha,\mu,\mathbf{k}|H|\alpha',\mu',\mathbf{k}\rangle &= \sum_{\mathbf{R}'\in\Omega}\langle\alpha,\boldsymbol{\tau}_\mu|H|\alpha',\mathbf{R}'+\boldsymbol{\tau}_{\mu'}\rangle \\ &\quad\times e^{i\mathbf{k}\cdot(\mathbf{R}'+\boldsymbol{\tau}_{\mu'}-\boldsymbol{\tau}_\mu)}.\end{aligned} \tag{5.28}$$

The kinematic momentum operator (4.24) is related to this Hamiltonian by

$$\begin{aligned}\langle\alpha,\mu,\mathbf{k}|\mathbf{p}|\alpha',\mu',\mathbf{k}\rangle &= \frac{m}{i\hbar}(\mathbf{x}_\alpha-\mathbf{x}_{\alpha'}+i\nabla_{\mathbf{k}}) \\ &\quad\times\langle\alpha,\mu,\mathbf{k}|H|\alpha',\mu',\mathbf{k}\rangle,\end{aligned} \tag{5.29}$$

where the original basis was assumed to be given by (5.20). Note that (5.29) has the form of an intra-atomic matrix element (proportional to $\mathbf{x}_\alpha-\mathbf{x}_{\alpha'}$) plus an inter-atomic matrix element (proportional to $i\nabla_{\mathbf{k}}$). The zero-parameter model of Eq. (1.3) is obtained in the limit $\mathbf{x}_\alpha\to 0$.

Let us now consider specific examples of H for tetrahedral semiconductors. The simplest tight-binding model for the diamond or zinc-blende structure consists of a single s orbital per atom, which is obtained by putting one coordinate eigenket at each atomic position [see Eq. (2.13)]. In this model, as shown in Appendix C, each atom is linked to 4 nearest neighbors and 12 second-nearest neighbors. More distant linkages do not exist because the Voronoi polyhedra for these atoms do not touch one another.

Such a simple model is, of course, unable to describe even the qualitative features of tetrahedral semiconductors. The simplest conventional tight-binding model that works in this case is the sp^3 model.[2,3,4,19] The basic features of this model are described in Table I, which lists the basis size (number of orbitals per atom) and the number of independent Hamiltonian matrix elements for coupling between atoms in the diamond structure out to third nearest neighbors.[19] Table I also lists the properties of other tight-binding models used in the literature, such as sp^3s^*,[68] sp^3d^2,[11] and $sp^3d^5s^*$.[51] The number of parameters listed in this table is the number permitted by the symmetry of the model, which is not necessarily the same as that used in any specific implementation in the literature.

TABLE I: Number of free parameters for different tight-binding models in the diamond structure. The upper half of the table refers to conventional tight-binding models, while the lower half refers to models obtained from symmetrized coordinate eigenkets.

Model		Parameters			
Basis	Size	On-site	1NN	2NN	3NN
sp^3	4	2	4	7	7
sp^3s^*	5	4	7	11	11
sp^3d^2	6	3	7	13	13
$sp^3d^5s^*$	10	7	17	33	33
$\|111\rangle$ (T_d)	4	2	1	1	0
$\|100\rangle$	6	2	1	1	0
$\|111\rangle$ (O_h)	8	3	3	1	0
$\|110\rangle$	12	3	1	1	0
$\|110\rangle+\|000\rangle$	13	5	1	1	0
$\|111\rangle+\|100\rangle$	14	7	5	1	0
$\|111\rangle+\|100\rangle+\|000\rangle$	15	11	5	1	0

For comparison, the bottom half of Table I lists the properties of several tight-binding models constructed from symmetrized coordinate eigenkets. The number of free parameters for the models generated by $|100\rangle$, $|111\rangle$, and $|110\rangle$ may be deduced easily from the link geometry results presented in Appendix C. The corresponding numbers for the compound models (with more than one generator) were determined using the algorithm in Appendix B.

The most striking feature of Table I is the relative paucity of free parameters in the symmetrized coordinate approach, which occurs because of the restriction (4.16) imposed by local gauge symmetry. Several of the symmetrized coordinate models are direct analogs of conventional models (i.e., they have identical symmetry); for example, the $|111\rangle$ (T_d) model corresponds to sp^3, and the $|100\rangle$ model corresponds to sp^3d^2 [see Eqs. (2.10) and (2.13)]. However, the number of free parameters in conventional tight-binding theory grows steadily with distance, whereas in the present theory there is no coupling beyond second nearest neighbors.

This dearth of adjustable parameters means that the smallest basis sets do not provide a reliable model for the energy band structure. For example, in the $|111\rangle$ (T_d) model, the splitting of the bonding and antibonding s states at the Γ point is the same as that of the p states

at Γ (and that of the p states at X). In the $|100\rangle$ model, there is no coupling at all between p orbitals on different atoms at the Γ point, so the splitting of bonding and antibonding states is zero. Such difficulties arise primarily because there is only one nearest-neighbor coupling parameter in these models.

To increase the number of adjustable parameters without an undue increase in basis size, one must deliberately search for models with the *most complicated* topology available. A good starting point is the $|111\rangle$ (O_h) basis, which already has 3 nearest-neighbor parameters. Combining this with a $|100\rangle$ basis raises that number to 4 or 5, depending on the relative values of the coordinates in the two generators. This 14-orbital model is a dramatic improvement over any of the smaller models; however, an extra $|000\rangle$ site adds substantial extra flexibility without much change in the basis size. Thus, the 15-orbital model generated by $|000\rangle$, $|111\rangle$, and $|100\rangle$ is probably the smallest basis capable of describing tetrahedral semiconductors accurately. In the language of conventional tight-binding theory, this would be referred to as an $s^3p^6d^5f$ model.

As shown in Table I, this model has 17 free parameters (one of which is just the reference energy). Specific definitions for these parameters are given in Table II, which also presents parameter values for Ge and Si obtained by fitting the band structure of the 15-orbital model to that given by the nonlocal empirical pseudopotentials of Chelikowsky and Cohen.[69,70,71] Local Hamiltonian matrix elements are labeled V, whereas nonlocal on-site, nearest-neighbor, and second-nearest-neighbor terms are denoted α, β, and γ, respectively. The subscripts a, b, e, and f refer to independent sites generated by $|0,0,0\rangle$, $|r,r,r\rangle$, $|-r,-r,-r\rangle$, and $|r',0,0\rangle$, respectively. The labels a, e, and f are the Wyckoff labels for sites of different symmetry in the diamond structure.[52] The label b is not correct Wyckoff notation; it actually refers to an independent e site, but the notation b was used here because these sites lie on the bonds between atoms.

The energy band structure calculated from the parameters in Table II is plotted in Fig. 2. One can see that the 15-orbital model provides a good fit to the nonlocal pseudopotential bands from the bottom of the valence band to about 5 eV above the top of the valence band. Qualitative errors begin to occur near 9 eV at both the L and X points. For example, the X_1 level near 9 eV in both figures should occur above 12 eV. This discrepancy can be eliminated, but the author has not found any way of doing so without adversely affecting the quality of the overall fit.

It should be emphasized that the parameters in Table II are presented here merely as "proof of concept;" they are in no way intended as the final word on the subject, and the author would be surprised if a better set were not found in the future. The quantities included in the fitting procedure were the valence- and conduction-band energy levels at Γ, X, L, and K. Effective masses and deformation potentials were not considered, and no attempt was made at ensuring transferability.

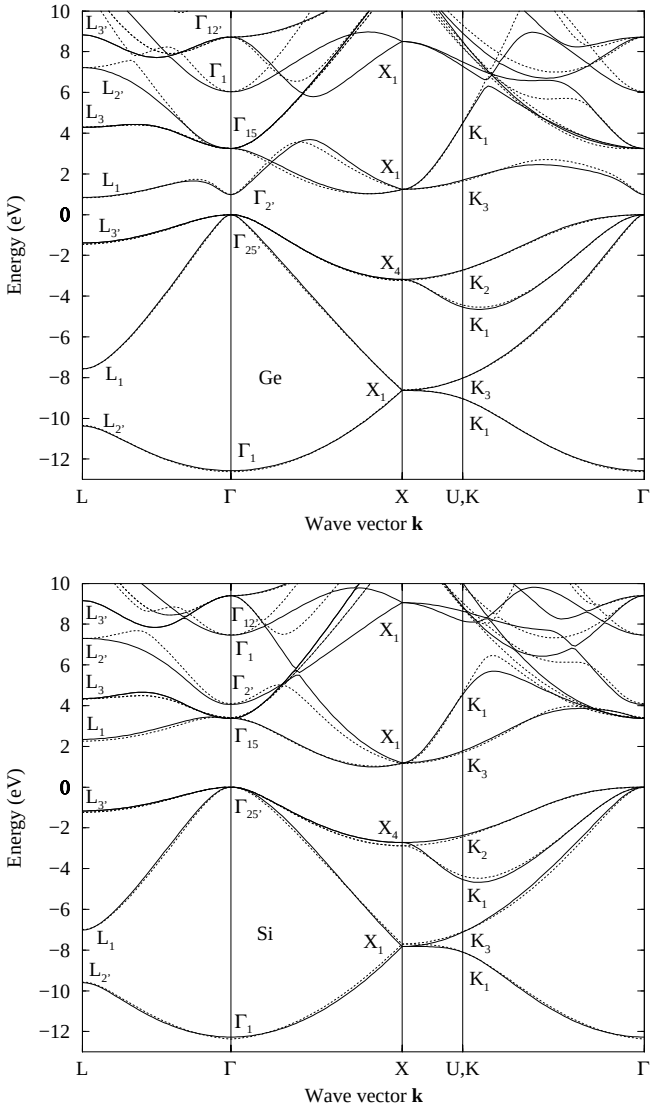

FIG. 2: Energy band structure of germanium and silicon. Solid lines: 15-orbital tight-binding model based on parameters in Table II. Dotted lines: Nonlocal empirical pseudopotential model of Ref. 71. The kinetic-energy cutoff for the latter calculation was $(\mathbf{k}+\mathbf{G})^2 \le 21(2\pi/a_0)^2$, which corresponds to 113 plane waves at Γ.

The main difficulty encountered during the fitting was the lack of any reliable method for establishing a sound starting point. Unlike the case for smaller tight-binding models, the present Hamiltonian has almost no simple analytical solutions (except for the $\Gamma_{12'}$, Γ_{12}, and X_2 states, which are relatively unimportant) that can be used to determine starting values. The formula (4.42) for the kinetic energy provides a set of "free-particle" parameters that is better than nothing, but in a 15-orbital basis, Eq. (4.42) is a poor approximation to the continuum kinetic energy (5.19). After several months and dozens of different schemes (which still sampled only an infinitesi-

TABLE II: Independent Hamiltonian parameters in the 15-orbital model generated by $|0,0,0\rangle$, $|r,r,r\rangle$, and $|r',0,0\rangle$.

Parameter		Value (Ry)	
Symbol	Definition	Ge	Si
V_a	$\langle 0,0,0\|H\|0,0,0\rangle$	1.22536	1.76282
V_b	$\langle r,r,r\|H\|r,r,r\rangle$	1.18998	1.65167
V_e	$\langle -r,r,r\|H\|-r,r,r\rangle$	1.07211	1.17674
V_f	$\langle r',0,0\|H\|r',0,0\rangle$	1.77902	1.98485
α_{ab}	$-\langle 0,0,0\|H\|r,r,r\rangle$	−0.56483	−0.50749
α_{ae}	$-\langle 0,0,0\|H\|-r,r,r\rangle$	0.07052	0.24801
α_{af}	$-\langle 0,0,0\|H\|r',0,0\rangle$	0.06400	−0.07980
α_{be}	$-\langle r,r,r\|H\|-r,r,r\rangle$	0.03841	0.19402
α_{bf}	$-\langle r,r,r\|H\|r',0,0\rangle$	0.69832	0.78464
α_{ef}	$-\langle -r,r,r\|H\|0,r',0\rangle$	0.31958	0.13684
α_{ff}	$-\langle r',0,0\|H\|0,r',0\rangle$	−0.43768	−0.47311
β_{bb}	$-\langle r,r,r\|H\|a-r,a-r,a-r\rangle$	0.74306	1.45766
β_{ee}	$-\langle -r,r,r\|H\|a-r,a-r,a+r\rangle$	−0.03011	0.01537
β_{bf}	$-\langle r,r,r\|H\|a-r',a,a\rangle$	−0.07041	−0.35201
β_{ef}	$-\langle -r,r,r\|H\|a-r',a,a\rangle$	0.28640	0.40040
β_{ff}	$-\langle r',0,0\|H\|a,a-r',a\rangle$	0.13140	0.17401
γ_{ee}	$-\langle -r,r,r\|H\|-r,2a-r,2a-r\rangle$	0.02976	0.02025

mal fraction of parameter space), the author was unable to find any method whose success could honestly be attributed to anything other than trial and error. Hence, the development of a robust fitting procedure remains an unsolved problem.

C. Dielectric function

As a test of the field-particle coupling in the 15-orbital model, the imaginary part of the transverse dielectric tensor was calculated from the formula[23,72]

$$\epsilon_2^{\alpha\beta}(\omega) = \frac{4\pi^2e^2}{m^2\omega^2}\sum_{c,v}\frac{2}{(2\pi)^3}\int_{\Omega_0^*}\langle v\mathbf{k}|p^\alpha|c\mathbf{k}\rangle\langle c\mathbf{k}|p^\beta|v\mathbf{k}\rangle \times \delta(E_{c\mathbf{k}}-E_{v\mathbf{k}}-\hbar\omega)\mathrm{d}^3k, \quad (5.30)$$

where $\hbar\omega$ is the photon energy, and $|n\mathbf{k}\rangle$ is an eigenket of H with energy $E_{n\mathbf{k}}$. The sum covered the four valence bands v and the seven lowest conduction bands c. The integral was performed using a modified Gilat-Raubenheimer technique[73,74,75] based on 45961 $\mathbf{k}$ points in the irreducible part of the Brillouin zone[76,77] (representing 2048000 points in the full Brillouin zone). The energy interval for this calculation was 1 meV.

To reveal more clearly the physical meaning of the calculated spectra, the same method was used to calculate the joint density of states function

$$J(E)=\sum_{c,v}\frac{2\Omega_0}{(2\pi)^3}\int_{\Omega_0^*}\delta(E_{c\mathbf{k}}-E_{v\mathbf{k}}-E)\mathrm{d}^3k. \quad (5.31)$$

The dielectric function differs from $J(E)$ in that each transition is weighted by the oscillator strength

$$f_{cv}^{\alpha\beta}=\frac{2\langle v\mathbf{k}|p^\alpha|c\mathbf{k}\rangle\langle c\mathbf{k}|p^\beta|v\mathbf{k}\rangle}{m(E_{c\mathbf{k}}-E_{v\mathbf{k}})}. \quad (5.32)$$

One may therefore use ϵ_2 and J to define an average oscillator strength at each energy by

$$F^{\alpha\beta}(\hbar\omega)=\frac{m\omega\Omega_0}{2\hbar\pi^2e^2}\frac{\epsilon_2^{\alpha\beta}(\omega)}{J(\hbar\omega)}. \quad (5.33)$$

In cubic crystals, the tensors (5.30) and (5.33) reduce to scalars: $\epsilon_2^{\alpha\beta}(\omega)=\epsilon_2(\omega)\delta_{\alpha\beta}$.

The calculated dielectric function $\epsilon_2(\omega)$ for Ge and Si is plotted in Fig. 3. This figure compares experimental data[78] for the dielectric function with the values given by Eq. (5.30) for (i) the nonlocal pseudopotential model of Chelikowsky and Cohen,[69,70,71] and (ii) the 15-orbital tight-binding model of Table II. For each model, two plots of $\epsilon_2(\omega)$ are given, corresponding to two different expressions for the momentum operator $\mathbf{p}$.

In pseudopotential calculations, optical properties are usually calculated from $\mathbf{A}\cdot\mathbf{p}$ coupling with $\mathbf{p}=-i\hbar\nabla$.[71] However, if the pseudopotential is nonlocal, this coupling is not gauge invariant; the correct linear coupling is given instead by the kinematic momentum $\mathbf{p}=(m/i\hbar)[\mathbf{x},H]$.[9] Since pseudopotential calculations are usually performed in a plane-wave basis, a more convenient expression for the kinematic momentum is given by Eq. (5.9) (which is valid for both discrete and continuous coordinates $\mathbf{x}$).

In the present tight-binding theory, the kinematic momentum is given by Eq. (5.29). The two tight-binding functions plotted in Fig. 3 correspond to two different choices of the intra-atomic coordinates $\mathbf{x}_\alpha$, which are determined by the parameters r and r' in Table II. One choice was the limit $r\to 0$, $r'\to 0$, which is equivalent to the zero-parameter model of Eqs. (1.3) and (1.4). The other, more physically realistic choice was $r=\frac{1}{3}a$ and $r'=\sqrt{2}r$. The value $r=\frac{1}{3}a$ was chosen because it yields equidistant lattice sites along the bond directions $\langle 111\rangle$. The value $r'=\sqrt{2}r$ was used because a somewhat larger

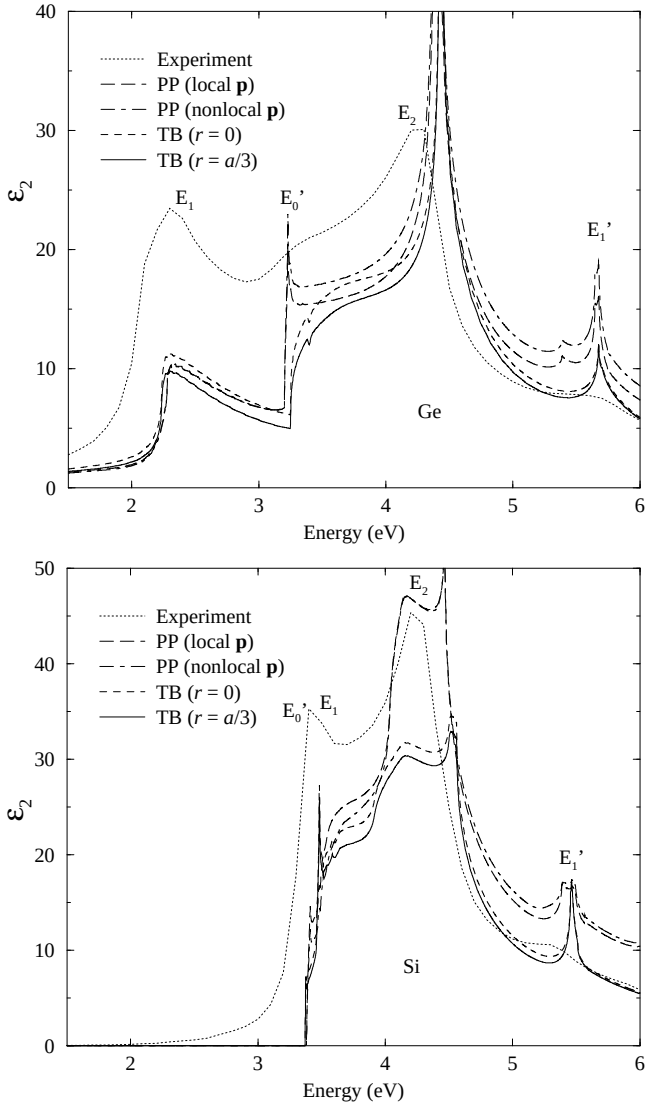

FIG. 3: Imaginary part of the transverse dielectric function of germanium and silicon. Dotted line: Experimental data from Ref. 78. Long dashed line: Nonlocal pseudopotential model of Ref. 71 with canonical (local) momentum $\mathbf{p} = -i\hbar\nabla$. Dot-dashed line: Nonlocal pseudopotential model of Ref. 71 with kinematic (nonlocal) momentum $\mathbf{p} = (m/i\hbar)[\mathbf{x}, H]$. Short dashed line: 15-orbital tight-binding model from Table II with $r \to 0$ and $r' \to 0$. Solid line: 15-orbital tight-binding model from Table II with $r = \frac{1}{3}a$ and $r' = \sqrt{2}r$.

value (e.g., $1.5r$) breaks the link α_{ff} in Table II, whereas a somewhat smaller value (e.g., $1.3r$) breaks the links β_{bf} and β_{ff} (while simultaneously forming a new link β_{be}). These values of r and r' also generated successful starting values for some of the parameters in Table II (although the final fitted parameters were not very close to the starting values).

Several conclusions may be drawn from Fig. 3. The first is that, within a given model (pseudopotential or tight binding), the choice of momentum operator does not have much numerical significance for the present calculation. This was to be expected on physical grounds, since the intra-atomic coordinate $\mathbf{x}_\alpha$ (in the tight-binding model) and the nonlocal part of the momentum (in the pseudopotential model) both lead to polarization effects *within* the atom. Yet it is well known that the bonds between atoms are much easier to polarize than the atoms themselves.[3,4] Hence, in a bulk semiconductor, intra-atomic effects yield only a minor numerical correction. This conclusion should remain valid in any system where the states are extended, but it may break down in systems where localized states are important.[32,33]

The nonlocal part of the pseudopotential momentum tends to increase $\epsilon_2(\omega)$ in most frequency ranges, but it sometimes has the opposite effect (see, e.g., the region below 4 eV in Si). However, the intra-atomic coupling in the tight-binding model always decreases $\epsilon_2(\omega)$. This may be understood by noting that the dominant nonlocal term in the Hamiltonian of Table II is the coupling β_{bb} along the bond between nearest neighbors. Increasing the value of r decreases the distance between b sites on neighboring atoms, thereby decreasing the momentum matrix in Eq. (5.29). This tends to increase (slightly) the discrepancy between the tight-binding and pseudopotential dielectric functions. It is possible, however, that a different parametrization of the Hamiltonian might yield different results.

The tight-binding and pseudopotential dielectric functions are quite similar in Ge, but there is a significant discrepancy at the E_2 peak in Si. The reason for the difference between the models is apparent from the joint density of states J and average oscillator strength F plotted in Fig. 4. This figure shows that the tight-binding J is very accurate in Ge and somewhat less so in Si, as might have been expected from the quality of the fitted energy bands in Fig. 2. However, the tight-binding model underestimates the oscillator strength over almost the entire frequency range shown, typically by about 20%. The difference is most pronounced between 4 and 4.5 eV in Si, where it exceeds 30%. When combined with a slight underestimate of J in the same region, this leads to the discrepancy in the E_2 peak noted above.

The E_2 peak in Si is associated with a volume in $\mathbf{k}$ space near $(0.9, 0.1, 0.1)2\pi/a_0$,[71] which is close to the X point. Thus, the physical reason for the error in the E_2 peak is probably the spurious X_1 conduction band near 9 eV in Fig. 2. Because this band is too low in energy, it mixes more strongly with the lowest X_1 conduction band in the tight-binding model than it does in the pseudopotential model. This change in the wave function causes a corresponding change in oscillator strength. Thus, the majority of the error in the Si E_2 peak would likely be eliminated if one could find an improved parameter set that raises the energy of the upper X_1 conduction band.

It is less clear whether the systematic underestimate of oscillator strength at all frequencies could be resolved by changing the Hamiltonian parameters. Oscillator

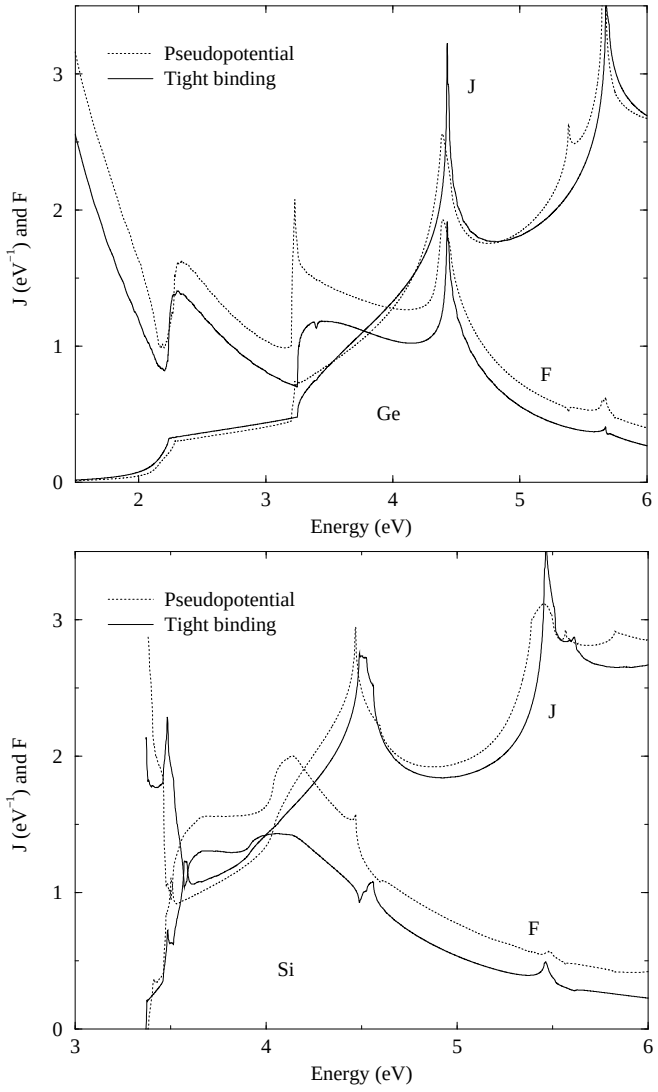

FIG. 4: Joint density of states J and average oscillator strength F for Ge and Si. Dotted line: Nonlocal pseudopotential model of Ref. 71 with kinematic momentum $\mathbf{p} = (m/i\hbar)[\mathbf{x}, H]$. Solid line: 15-orbital tight-binding model from Table II with $r = \frac{1}{3}a$ and $r' = \sqrt{2}r$.

strength was not included in the present fitting routine, so it is possible that with specific attention to this feature, one could improve the oscillator strength while maintaining the quality of the joint density of states. However, it is also possible that such an underestimate is a fundamental limitation imposed by the small basis size in the tight-binding model. Thus, at present, the 15-orbital model is capable of providing semiquantitative predictions of oscillator strength that reproduce all of the major trends exhibited by the pseudopotential model. Whether future developments bring it into precise quantitative agreement remains to be seen.

Finally, it is worth noting that in Fig. 3, the calculated E_1 peak (between 2 and 3 eV) for Ge is considerably less than the experimental value for both the pseudopotential and tight-binding models. Chelikowsky and Cohen have attributed this discrepancy to the neglect of exciton effects.[69] However, a recent first-principles calculation of $\epsilon_2(\omega)$ that includes electron-hole interactions[79] shows that the contribution from excitons in Ge is not large enough to fill the gap in Fig. 3. Thus, the empirical pseudopotential for Ge[69,71] probably needs further adjustment to increase the E_1 peak.

VI. DISCUSSION AND CONCLUSIONS

This paper has shown that intra-atomic optical transitions can be incorporated into tight-binding theory in a gauge-invariant way if the coordinate representation is taken as fundamental. Orthogonal atomic-like orbitals can be constructed from symmetrized coordinate eigenkets, and the coupling to electromagnetic fields can then be described using lattice gauge theory. A model based on 15 such orbitals per atom is capable of describing the most important features of the tetrahedral semiconductors Ge and Si. This basis is slightly larger than existing 10-orbital models,[51] but it has the advantages of (i) gauge invariance and (ii) providing an explicit wave function for the electron. A larger basis is needed in the present theory because the restrictions imposed by local gauge symmetry reduce the number of available Hamiltonian fitting parameters.

The field-particle coupling derived here is similar to that given by the Peierls substitution,[9,22,23,24,25,26] but the field-induced phase factor appears in the coordinate representation rather than the tight-binding representation. Thus, the present formalism includes intra-atomic coupling not present in the Peierls substitution. Ismail-Beigi, Chang, and Louie[9] have recently presented a derivation of the Peierls phase for nonlocal Hamiltonians in a continuous coordinate representation. They have argued that this derivation justifies the use of the Peierls substitution in tight-binding theory. However, their derivation cannot be extrapolated to tight-binding theory, because ordinary $\mathbf{A}\cdot\mathbf{p}$ coupling gives rise to intra-atomic interactions that are not included in the Peierls substitution. The existence of such interactions was not considered in the tight-binding theory of Ref. 9.

It is interesting to consider whether there are any other ways of incorporating local gauge symmetry into tight-binding theory. One possibility is to work directly in the usual tight-binding representation [see, e.g., Eq. (1.4)], where the basis kets are labeled by the symmetry of the orbital and the position of the atom. If gauge symmetry is to be applied in this basis, the coordinate operator must be diagonal, hence all intra-atomic matrix elements must be set to zero [as in Eq. (1.3)]. One could then introduce an Abelian $U(1)$ gauge field on this lattice using the approach described above in Sec. IV. The results would be identical to those found in Sec. IV, except that the phase factor in the Hamiltonian (4.22) would be applied in the tight-binding representation rather than the

coordinate representation. Hence, this approach would constitute a "derivation" of the Peierls phase (1.4). Such a derivation would eliminate the ambiguity associated with the choice of path[23] in Eq. (1.4). (Other techniques for eliminating path ambiguity are described in Refs. 9 and 25.)

The problem with this approach lies in its treatment of the coordinate operator. In the theory described in Sec. IV, when the basis size is increased, the eigenvalue spectrum of the coordinate operator remains nondegenerate, tending (in the limit of infinite basis dimensions) toward a continuous spectrum. However, if the coordinate operator is required to be diagonal in the tight-binding basis, its eigenvalue spectrum is always degenerate, tending (in the limit of infinite basis dimensions) toward a discrete spectrum with infinite degeneracy. Hence, any tight-binding theory that is either based on or equivalent to the Peierls substitution cannot reproduce the correct continuum limit of the coordinate operator.

As a generalization of the above approach, one might also consider introducing a non-Abelian gauge field[37,38,39,40,43,44,45,46,47,48,49] in the tight-binding basis. The idea would be to treat the tight-binding electron as a new type of "elementary particle" with some internal degrees of freedom (corresponding to the symmetry labels of the atomic orbitals) that are coupled to the gauge field. In this way, one might hope to reproduce the effects of intra-atomic coupling while remaining in the tight-binding basis. There are, however, numerous difficulties with this approach.

First, the lattice sites in lattice gauge theory represent states of the same particle at different positions. Hence, these states are identical apart from their positions. However, in tight-binding theory the atoms are generally not the same. Second, the field equations for a non-Abelian gauge field are intrinsically nonlinear, because the field carries its own charge and is coupled directly to itself (i.e., it is self-radiating). It is therefore difficult to imagine how such a field could reproduce ordinary electromagnetism, in which the field has no charge, and nonlinearities arise only from interactions with matter. Third, one would need to define a new gauge field theory every time one added new orbitals to the model, and every one of these non-Abelian field theories would need to reproduce the results of Abelian electromagnetic theory. Finally, the coordinate operator in this approach would still have a discrete, degenerate eigenvalue spectrum.

Thus, it appears that the present approach—that is, an Abelian $U(1)$ gauge field in the coordinate representation—is the only gauge-invariant method for including electromagnetic fields in empirical tight-binding theory that tends toward the correct continuum limit as the basis dimensions are increased. In this case, the only way that the essential structure of the theory can be modified is to change the topology of the system so as to increase the number of links between lattice sites. This would increase the number of free parameters in the Hamiltonian, thereby permitting a reduction in basis size. Such a modification would clearly be beneficial, but it is not obvious that there exists any alternative topology for general lattices that is capable of reproducing continuum electromagnetism unambiguously. Hence, this possibility will not be explored further here.

Acknowledgments

I am grateful to Peter Vogl, Tim Boykin, and Tai Kai Ng for helpful discussions. This work was supported by Hong Kong RGC Grant No. HKUST6139/00P.

APPENDIX A: SYMMETRIZED ORBITALS

This section presents symmetrized orbitals obtained by applying the symmetry operations of the cubic group O_h to the coordinate eigenkets $|100\rangle$ and $|110\rangle$. [The orbitals for $|111\rangle$ may be found in Eq. (2.11).] For $|100\rangle$, if the basis kets are ordered as $\{|100\rangle, |010\rangle, |001\rangle, |\bar{1}00\rangle, |0\bar{1}0\rangle, |00\bar{1}\rangle\}$, then the symmetrized orbitals are

$$\begin{aligned}
|\Gamma_1\rangle &= |s\rangle \\
&= \tfrac{1}{\sqrt{6}}(1,1,1,1,1,1), \\
|\Gamma_{15}^z\rangle &= |p_z\rangle \\
&= \tfrac{1}{\sqrt{2}}(0,0,1,0,0,-1), \\
|\Gamma_{12}^a\rangle &= |d_{2z^2-x^2-y^2}\rangle \\
&= \tfrac{1}{2\sqrt{3}}(-1,-1,2,-1,-1,2), \\
|\Gamma_{12}^b\rangle &= |d_{x^2-y^2}\rangle \\
&= \tfrac{1}{2}(1,-1,0,1,-1,0).
\end{aligned} \tag{A1}$$

For $|110\rangle$, if the basis kets are ordered as $\{|011\rangle, |01\bar{1}\rangle, |0\bar{1}1\rangle, |0\bar{1}\bar{1}\rangle, |101\rangle, |10\bar{1}\rangle, |\bar{1}01\rangle, |\bar{1}0\bar{1}\rangle, |110\rangle, |1\bar{1}0\rangle, |\bar{1}10\rangle, |\bar{1}\bar{1}0\rangle\}$, then the symmetrized orbitals are

$$\begin{aligned}
|\Gamma_1\rangle &= |s\rangle \\
&= \tfrac{1}{2\sqrt{3}}(1,1,1,1,1,1,1,1,1,1,1,1), \\
|\Gamma_{15}^z\rangle &= |p_z\rangle \\
&= \tfrac{1}{2\sqrt{2}}(1,-1,1,-1,1,-1,1,-1,0,0,0,0), \\
|\Gamma_{12}^a\rangle &= |d_{2z^2-x^2-y^2}\rangle \\
&= \tfrac{1}{2\sqrt{6}}(-1,-1,-1,-1,-1,-1,-1,-1,2,2,2,2), \\
|\Gamma_{12}^b\rangle &= |d_{x^2-y^2}\rangle \\
&= \tfrac{1}{2\sqrt{2}}(1,1,1,1,-1,-1,-1,-1,0,0,0,0), \\
|\Gamma_{25'}^{xy}\rangle &= |d_{xy}\rangle \\
&= \tfrac{1}{2}(0,0,0,0,0,0,0,0,1,-1,-1,1), \\
|\Gamma_{25}^c\rangle &= |f_{z(x^2-y^2)}\rangle \\
&= \tfrac{1}{2\sqrt{2}}(-1,1,-1,1,1,-1,1,-1,0,0,0,0).
\end{aligned} \tag{A2}$$

For the triply degenerate representations Γ_{15}, $\Gamma_{25'}$, and Γ_{25}, only one representative orbital is given; the others may be obtained from cyclic permutations of x, y, and z.

APPENDIX B: GEOMETRY OF VORONOI POLYHEDRA

This appendix presents an algorithm for calculating the geometry of the Voronoi polyhedra associated with a given set of nodes $\mathbf{x}_i$. The basic element in this algorithm is a procedure for finding the edges of the polyhedra. An edge of a Voronoi polyhedron is a finite line segment consisting of points that are closer to three (or more) nodes than to any other nodes. The first step is therefore to determine the equation defining this line.

Any three noncollinear points $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$ define a plane whose normal is the vector

$$\begin{aligned}\mathbf{n}_{ijk} &= \mathbf{d}_{ji} \times \mathbf{d}_{ki} \\ &= \mathbf{x}_i \times \mathbf{x}_j + \mathbf{x}_j \times \mathbf{x}_k + \mathbf{x}_k \times \mathbf{x}_i, \end{aligned} \tag{B1}$$

where $\mathbf{d}_{ji} = \mathbf{x}_j - \mathbf{x}_i$. This plane is the set of points $\mathbf{x}$ satisfying

$$\mathbf{n}_{ijk} \cdot (\mathbf{x} - \mathbf{x}_i) = 0. \tag{B2}$$

The line consisting of all points $\mathbf{x}$ equidistant from $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$ may therefore be written as

$$\mathbf{x} = \mathbf{x}_{ijk} + \lambda \hat{\mathbf{n}}_{ijk}, \tag{B3}$$

where λ is a real parameter, $\hat{\mathbf{n}}_{ijk} = \mathbf{n}_{ijk}/n_{ijk}$, and $\mathbf{x}_{ijk}$ is the point in the plane (B2) equidistant from $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$. To determine this point, note that points $\mathbf{x}$ equidistant from $\mathbf{x}_i$ and $\mathbf{x}_j$ satisfy

$$(\mathbf{x}_j - \mathbf{x}_i) \cdot [\mathbf{x} - \tfrac{1}{2}(\mathbf{x}_j + \mathbf{x}_i)] = 0. \tag{B4}$$

The point $\mathbf{x}_{ijk}$ therefore satisfies the three equations

$$\mathbf{a}_r \cdot \mathbf{x}_{ijk} = c_r \qquad (r = 1, 2, 3) \tag{B5}$$

in which

$$\mathbf{a}_1 = \mathbf{d}_{ji}, \qquad \mathbf{a}_2 = \mathbf{d}_{ki}, \qquad \mathbf{a}_3 = \mathbf{n}_{ijk}, \tag{B6}$$

which may be viewed as a set of oblique (more specifically, monoclinic) basis vectors, and

$$\begin{aligned} c_1 &= \tfrac{1}{2}(x_j^2 - x_i^2), \\ c_2 &= \tfrac{1}{2}(x_k^2 - x_i^2), \\ c_3 &= \mathbf{n}_{ijk} \cdot \mathbf{x}_i \\ &= \mathbf{x}_i \cdot (\mathbf{x}_j \times \mathbf{x}_k). \end{aligned} \tag{B7}$$

The solution to Eqs. (B5) is given by

$$\mathbf{x}_{ijk} = \sum_{s=1}^{3} c_s \mathbf{b}_s, \tag{B8}$$

where $\mathbf{b}_s$ is a reciprocal basis vector satisfying $\mathbf{a}_r \cdot \mathbf{b}_s = \delta_{rs}$; e.g.,

$$\mathbf{b}_1 = \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}. \tag{B9}$$

Now since $\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3) = n_{ijk}^2$, the vectors $\mathbf{b}_s$ are given explicitly by

$$\begin{aligned} \mathbf{b}_1 &= [\mathbf{d}_{ji}(d_{ki}^2) - \mathbf{d}_{ki}(\mathbf{d}_{ji} \cdot \mathbf{d}_{ki})]/n_{ijk}^2, \\ \mathbf{b}_2 &= [\mathbf{d}_{ki}(d_{ji}^2) - \mathbf{d}_{ji}(\mathbf{d}_{ji} \cdot \mathbf{d}_{ki})]/n_{ijk}^2, \\ \mathbf{b}_3 &= \mathbf{n}_{ijk}/n_{ijk}^2. \end{aligned} \tag{B10}$$

Equation (B8) may then be rearranged in the more symmetric form

$$\mathbf{x}_{ijk} = \frac{\mathbf{x}_i[d_{jk}^2(\mathbf{d}_{ji} \cdot \mathbf{d}_{ki})] + \mathbf{x}_j[d_{ki}^2(\mathbf{d}_{kj} \cdot \mathbf{d}_{ij})] + \mathbf{x}_k[d_{ij}^2(\mathbf{d}_{ik} \cdot \mathbf{d}_{jk})]}{2n_{ijk}^2}. \tag{B11}$$

This result, together with Eq. (B3), defines the line equidistant from nodes $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$.

The next step is to determine whether any segment of this line forms an edge of a Voronoi polyhedron. Points on such a segment must lie closer to $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$ than to any other node $\mathbf{x}_l$. For each node $\mathbf{x}_l$, one calculates

$$\alpha_l = \mathbf{d}_{il} \cdot \hat{\mathbf{n}}_{ijk}. \tag{B12}$$

If $\alpha_l = 0$, then $\mathbf{x}_l$ lies in the plane (B2). In this case, if $|\mathbf{x}_l - \mathbf{x}_{ijk}| < |\mathbf{x}_i - \mathbf{x}_{ijk}|$, then no portion of the line (B3) forms an edge of a Voronoi polyhedron. On the other hand, if $|\mathbf{x}_l - \mathbf{x}_{ijk}| \geq |\mathbf{x}_i - \mathbf{x}_{ijk}|$, then the line (B3) may form an edge (depending on the position of the other nodes $\mathbf{x}_{l'}$).

If $\alpha_l \neq 0$, then $\mathbf{x}_l$ does not lie in the plane (B2). In this case, points on the line (B3) that are closer to $\mathbf{x}_i$ than to $\mathbf{x}_l$ satisfy [cf. Eq. (B4)]

$$\mathbf{d}_{il} \cdot (\mathbf{x}_{ijk} + \lambda \hat{\mathbf{n}}_{ijk} - \mathbf{x}_{il}) > 0, \tag{B13}$$

in which $\mathbf{x}_{il} = \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_l)$. Hence, the position of the point equidistant from $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$, and $\mathbf{x}_l$ is given by the following value of the parameter λ in Eq. (B3):

$$\lambda_l = \frac{\mathbf{d}_{il} \cdot (\mathbf{x}_{il} - \mathbf{x}_{ijk})}{\alpha_l}. \tag{B14}$$

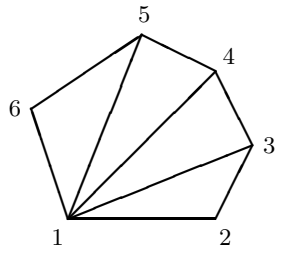

FIG. 5: Partitioning of the surface S_{ij} into triangles for the area calculation in Eq. (B20).

One may then define

$$\lambda_{\min} = \max(\lambda_l | \alpha_l > 0) \tag{B15}$$

(i.e., the maximum value of λ_l for all l such that $\alpha_l > 0$) and

$$\lambda_{\max} = \min(\lambda_l | \alpha_l < 0). \tag{B16}$$

If $\lambda_{\max} > \lambda_{\min}$, then the line segment (B3) with $\lambda_{\min} < \lambda < \lambda_{\max}$ forms an edge of a Voronoi polyhedron. This establishes the positions of two corners of the polyhedron:

$$\begin{aligned} \mathbf{x}_c &= \mathbf{x}_{ijk} + \lambda_{\min} \hat{\mathbf{n}}_{ijk}, \\ \mathbf{x}_{c'} &= \mathbf{x}_{ijk} + \lambda_{\max} \hat{\mathbf{n}}_{ijk}. \end{aligned} \tag{B17}$$

The set of all nodes in the plane (B2) that lie closer to the line segment $\lambda_{\min} < \lambda < \lambda_{\max}$ than any other node defines what is called a plaquette. Since there are in general more than three such nodes, it is convenient to define a unique label q for each plaquette, with

$$\hat{\mathbf{n}}_q \equiv \hat{\mathbf{n}}_{ijk}, \qquad \mathbf{x}_q \equiv \mathbf{x}_{ijk}, \tag{B18}$$

for any members $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$ of the given plaquette. (The sign of $\hat{\mathbf{n}}_q$ is fixed by some convention for the ordering of the nodes $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$.) Each plaquette is associated uniquely with one edge of a Voronoi polyhedron, the length of which is

$$d_q = |\mathbf{x}_{c'} - \mathbf{x}_c| = \lambda_{\max} - \lambda_{\min}, \tag{B19}$$

with $d_q > 0$ by definition.

At this point, one has sufficient information to determine whether a link exists between any pair of nodes $\mathbf{x}_i$ and $\mathbf{x}_j$. The first step is to use the above procedure to find all of the corner points $\mathbf{x}_c$ common to nodes i and j. By definition, all such points lie in the plane (B4). The set of these points defines a polygon, the perimeter of which consists of the line segments (B19). The area of the polygon may be calculated by numbering the corner points $\mathbf{x}_c$ in sequential order around the perimeter of the polygon, then partitioning the polygon into triangles as shown in Fig. 5. The normal area vector is

$$\mathbf{S}_{ij} = \frac{1}{2} \sum_{c=3}^{N_{ij}} (\mathbf{x}_{c-1} - \mathbf{x}_1) \times (\mathbf{x}_c - \mathbf{x}_1), \tag{B20}$$

where N_{ij} is the number of corner points common to nodes i and j. The area $S_{ij} = |\mathbf{S}_{ij}|$ is the area of the surface shared by the Voronoi polyhedra for sites i and j; note that if $N_{ij} = 1$ or 2, the shared region is a point or line, and the area (B20) is zero. Nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ are linked only if $S_{ij} > 0$.

The volume Ω_i of the Voronoi polyhedron for node $\mathbf{x}_i$ may be calculated from S_{ij} and d_{ij}. One simply integrates the identity $\nabla \cdot \mathbf{x} = 3$ over the polyhedron, using the divergence theorem and the fact that the plane containing S_{ij} is the perpendicular bisector of $\mathbf{d}_{ij}$ (although $\mathbf{d}_{ij}$ need not intersect S_{ij} itself). The result is

$$\Omega_i = \frac{1}{6} \sum_j S_{ij} d_{ij}. \tag{B21}$$

For each link $\ell \equiv (i,j)$, one can construct a polyhedron by drawing lines from nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ to each of their common corner points $\mathbf{x}_c$. The volume of this polyhedron is

$$\Omega_\ell = \frac{1}{3} S_\ell d_\ell. \tag{B22}$$

The volume $\Omega_\ell \equiv \Omega_{ij}$ is bisected by $S_\ell \equiv S_{ij}$, with half lying in Ω_i and half in Ω_j; hence

$$\Omega_i = \frac{1}{2} \sum_j \Omega_{ij}. \tag{B23}$$

The nodes in plaquette q define a polygon in the plane (B2); the perimeter of this polygon is formed by the links d_ℓ. Hence, the area of the plaquette can be calculated in the same way as the link area (B20):

$$\mathbf{S}_q = \frac{1}{2} \sum_{i=3}^{N_q} (\mathbf{x}_{i-1} - \mathbf{x}_1) \times (\mathbf{x}_i - \mathbf{x}_1), \tag{B24}$$

where N_q is the number of nodes in plaquette q. A polyhedron may be constructed for each plaquette by drawing lines from each node of the plaquette to the corner points (B17); the volume of this polyhedron is

$$\Omega_q = \frac{1}{3} S_q d_q. \tag{B25}$$

Finally, the plaquette surfaces S_q partition all of space into nonoverlapping polyhedra (this is referred to as a Delaunay tessellation[56]). These polyhedra (or *cells*) are in one-to-one correspondence with the corner points $\mathbf{x}_c$ of the Voronoi polyhedra. The volume of cell c is

$$\Omega_c = \frac{1}{3} \sum_{q \in c} \mathbf{S}_q \cdot (\mathbf{x}_q - \mathbf{x}_c), \tag{B26}$$

where the direction of $\mathbf{S}_q$ is chosen to point outward from Ω_c (note that $\mathbf{x}_c$ does not necessarily lie inside Ω_c,[47] so the dot product may be negative for some q).

A useful set of sum rules for verifying the consistency of a calculated geometry is

$$\sum_i \Omega_i = \sum_\ell \Omega_\ell = \sum_q \Omega_q = \sum_c \Omega_c = \Omega, \tag{B27}$$

where Ω is the volume of some region over which the node distribution is periodic, such as a primitive cell in a Bravais lattice (not to be confused with the generally nonperiodic cell Ω_c). The sum rule for Ω_i follows directly from the definition of Ω_i given in Sec. III, since every point in Ω must lie in at least one Voronoi polyhedron, and the only regions of overlap between polyhedra are points, lines, or planes of zero volume. The sum rule for Ω_c was proven in Ref. 47. The sum rule for Ω_ℓ follows from that for Ω_i, since the set $\{\Omega_\ell\}$ is just another way of partitioning the set $\{\Omega_i\}$ [see Eq. (B23)]. Likewise, the sum rule for Ω_q follows from that for Ω_c, since the set $\{\Omega_q\}$ is just another way of partitioning the set $\{\Omega_c\}$ [see Eqs. (B19), (B25), and (B26)].

APPENDIX C: EXAMPLES OF LINK GEOMETRY

This appendix presents values of the link lengths d_ℓ and surface areas S_ℓ for several lattices. The simplest geometry occurs for Bravais lattices, of which only the cubic lattices are considered here. For the simple cubic lattice, only nearest neighbors are linked, with $d_1 = a$ and $S_1 = a^2$, where a is the lattice constant. For the body-centered cubic lattice, both first and second nearest neighbors are linked, with

$$d_1 = \frac{\sqrt{3}}{2}a, \qquad S_1 = \frac{3\sqrt{3}}{16}a^2, \tag{C1}$$

and

$$d_2 = a, \qquad S_2 = \frac{1}{8}a^2. \tag{C2}$$

For the face-centered cubic lattice, only nearest neighbors are linked, with

$$d_1 = \frac{a}{\sqrt{2}}, \qquad S_1 = \frac{a^2}{4\sqrt{2}}. \tag{C3}$$

The remaining lattices to be considered are those obtained by putting symmetrized orbitals on the atomic sites of the diamond or zinc-blende structure. If only a single s orbital per atom is used (i.e., one $|000\rangle$ basis ket per atom), then each atom is linked to 4 nearest neighbors and 12 second-nearest neighbors, with

$$d_1 = \sqrt{3}a, \qquad S_1 = 3\sqrt{3}a^2, \tag{C4}$$

and

$$d_2 = 2\sqrt{2}a, \qquad S_2 = \frac{\sqrt{2}}{4}a^2. \tag{C5}$$

Here $a = \frac{1}{4}a_0$, where a_0 is the conventional cubic lattice constant. Note that in this case, the link d_2 does not intersect the surface S_2.

Coupling between second-nearest neighbors persists in models with more than one orbital per atom. In the "sp^3" model with four $|111\rangle$ sites per atom (generated by applying the symmetry operations of T_d to $|r,r,r\rangle$ and $|a-r,a-r,a-r\rangle$), each site is linked to three others on the same atom:

$$d_0 = 2\sqrt{2}r, \qquad S_0 = \frac{7\sqrt{2}}{4}a^2, \tag{C6}$$

one on a neighboring atom:

$$d_1 = \sqrt{3}(a-2r), \qquad S_1 = 3\sqrt{3}a^2, \tag{C7}$$

and three second-nearest neighbors:

$$d_2 = 2\sqrt{2}(a-r), \qquad S_2 = \frac{\sqrt{2}}{4}a^2. \tag{C8}$$

In the model generated by either T_d or O_h and $|r,0,0\rangle$, each of the six sites is linked to four others on the same atom:

$$d_0 = \sqrt{2}r, \qquad S_0 = \frac{a^2(5a-6r)}{4\sqrt{2}(a-r)}, \tag{C9}$$

four nearest neighbors:

$$d_1 = \sqrt{3a^2-4ar+2r^2}, \qquad S_1 = \frac{a^2 d_1}{2(a-r)}, \tag{C10}$$

and four second-nearest neighbors:

$$d_2 = \sqrt{2}(2a-r), \qquad S_2 = \frac{a^2(a-2r)}{4\sqrt{2}(a-r)}. \tag{C11}$$

In the model generated by either T_d or O_h and $|r,r,0\rangle$, each of the 12 sites is linked to four others on the same atom, two of which have

$$d_0 = \sqrt{2}r, \qquad S_0' = \frac{a^2(3a-4r)}{4\sqrt{2}(a-r)}, \tag{C12}$$

and two of which have

$$d_0 = \sqrt{2}r, \qquad S_0'' = \frac{a^2(7a-8r)}{4\sqrt{2}(a-r)}. \tag{C13}$$

Each site is also linked to two nearest neighbors:

$$d_1 = \sqrt{3a^2-8ar+6r^2}, \qquad S_1 = \frac{a^2 d_1}{2(a-r)}, \tag{C14}$$

and one second-nearest neighbor:

$$d_2 = 2\sqrt{2}(a-r), \qquad S_2 = \frac{a^3}{2\sqrt{2}(a-r)}. \tag{C15}$$

In the model generated by O_h and $|r,r,r\rangle$, there are two distinct lattice sites. Sites such as $|r,r,r\rangle$ and $|a-r,a-r,a-r\rangle$ are labeled b because they lie on the bonds between atoms, whereas sites such as $|-r,-r,-r\rangle$ and $|a+r,a+r,a+r\rangle$ are labeled e because they lie on "empty" bonds. (Both of these sites are actually Wyckoff e sites, but they are inequivalent, because the site symmetry of the atoms in diamond is T_d.) Each b site is linked to three e sites on the same atom (and vice versa), with

$$d_0^{be} = 2r, \qquad S_0^{be} = \frac{a(9a-4r)}{8}. \tag{C16}$$

Each b site is linked to one nearest-neighbor b site:

$$d_1^{bb} = \sqrt{3}(a-2r), \qquad S_1^{bb} = \frac{3\sqrt{3}}{4}a^2, \tag{C17}$$

and three nearest-neighbor e sites:

$$d_1^{be} = \sqrt{a^2+2(a-2r)^2}, \qquad S_1^{be} = \frac{1}{8}ad_1^{be}. \tag{C18}$$

Each e site is also linked to three b sites via (C18), to six nearest-neighbor e sites via

$$d_1^{ee} = \sqrt{2a^2+(a-2r)^2}, \qquad S_1^{ee} = \frac{1}{4}ad_1^{ee}, \tag{C19}$$

and to three second-nearest neighbor e sites via

$$d_2^{ee} = 2\sqrt{2}(a-r), \qquad S_2^{ee} = \frac{\sqrt{2}}{4}a(a+2r). \tag{C20}$$

* Electronic address: phbaf@ust.hk

[1] F. Bloch, Z. Phys. **52**, 555 (1928).
[2] J. C. Slater and G. F. Koster, Phys. Rev. **94**, 1498 (1954).
[3] W. A. Harrison, *Electronic Structure and the Properties of Solids* (Dover, New York, 1989).
[4] W. A. Harrison, *Elementary Electronic Structure* (World Scientific, Singapore, 1999).
[5] C. Z. Wang and K. M. Ho, Adv. Chem. Phys. **93**, 651 (1996).
[6] D. W. Bullett, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull (Academic, New York, 1980), vol. 35, pp. 129–214.
[7] C. M. Goringe, D. R. Bowler, and E. Hernández, Rep. Prog. Phys. **60**, 1447 (1997).
[8] P. E. A. Turchi, A. Gonis, and L. Colombo, eds., *Tight-Binding Approach to Computational Materials Science* (Materials Research Society, Warrendale, PA, 1998).
[9] S. Ismail-Beigi, E. K. Chang, and S. G. Louie, Phys. Rev. Lett. **87**, 087402 (2001).
[10] Y.-C. Chang and J. N. Schulman, Phys. Rev. B **31**, 2069 (1985).
[11] Y.-C. Chang and D. E. Aspnes, Phys. Rev. B **41**, 12002 (1990).
[12] Z. Xu, Solid State Commun. **76**, 1143 (1990).
[13] P.-O. Löwdin, J. Chem. Phys. **18**, 365 (1950).
[14] S. L. Altmann, *Band Theory of Solids: An Introduction from the Point of View of Symmetry* (Clarendon, Oxford, 1991), p. 222.
[15] G. Dresselhaus and M. S. Dresselhaus, Phys. Rev. **160**, 649 (1967).
[16] N. V. Smith, Phys. Rev. B **19**, 5019 (1979).
[17] L. Brey and C. Tejedor, Solid State Commun. **48**, 403 (1983).
[18] B. Koiller, R. Osório, and L. M. Falicov, Phys. Rev. B **43**, 4170 (1991).
[19] C. Tserbak, H. M. Polatoglou, and G. Theodorou, Phys. Rev. B **47**, 7104 (1993).
[20] L. C. Lew Yan Voon and L. R. Ram-Mohan, Phys. Rev. B **47**, 15500 (1993).
[21] T. B. Boykin, Phys. Rev. B **60**, 15810 (1999).
[22] R. Peierls, Z. Phys. **80**, 763 (1933).
[23] M. Graf and P. Vogl, Phys. Rev. B **51**, 4940 (1995).
[24] T. Dumitrică, J. S. Graves, and R. E. Allen, Phys. Rev. B **58**, 15340 (1998).
[25] T. B. Boykin, R. C. Bowen, and G. Klimeck, Phys. Rev. B **63**, 245314 (2001).
[26] T. B. Boykin and P. Vogl, Phys. Rev. B **65**, 035202 (2002).
[27] A. Selloni, P. Marsella, and R. Del Sole, Phys. Rev. B **33**, 8885 (1986).
[28] L. Reining, R. Del Sole, M. Cini, and J. G. Ping, Phys. Rev. B **50**, 8411 (1994).
[29] J. Bennetto and D. Vanderbilt, Phys. Rev. B **53**, 15417 (1996).
[30] K. Leung and K. B. Whaley, Phys. Rev. B **56**, 7455 (1997).
[31] K. Leung, S. Pokrant, and K. B. Whaley, Phys. Rev. B **57**, 12291 (1998).
[32] M. Cruz, M. R. Beltrán, C. Wang, J. Tagüeña-Martínez, and Y. G. Rubo, Phys. Rev. B **59**, 15381 (1999).
[33] T. G. Pedersen, K. Pedersen, and T. B. Kriestensen, Phys. Rev. B **63**, 201101 (2001).
[34] M. Governale and C. Ungarelli, Phys. Rev. B **58**, 7816 (1998).
[35] H. Weyl, Z. Phys. **56**, 330 (1929).
[36] P. A. M. Dirac, Proc. R. Soc. (London) A **133**, 60 (1931).
[37] C. N. Yang and R. L. Mills, Phys. Rev. **96**, 191 (1954).
[38] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).
[39] K. Moriyasu, *An Elementary Primer for Gauge Theory* (World Scientific, Singapore, 1983).
[40] M. Guidry, *Gauge Field Theories* (Wiley, New York, 1991).
[41] M. Tinkham, *Group Theory and Quantum Mechanics* (McGraw-Hill, New York, 1964).
[42] F. Bassani, in *Semiconductors and Semimetals*, edited by R. K. Willardson and A. C. Beer (Academic, New York, 1966), vol. 1, pp. 21–74.
[43] K. G. Wilson, Phys. Rev. D **10**, 2445 (1974).
[44] J. Kogut and L. Susskind, Phys. Rev. D **11**, 395 (1975).
[45] C. Rebbi, *Lattice Gauge Theories and Monte Carlo Simulations* (World Scientific, Singapore, 1983).
[46] H. J. Rothe, *Lattice Gauge Theories: An Introduction* (World Scientific, Singapore, 1997), 2nd ed.
[47] N. H. Christ, R. Friedberg, and T. D. Lee, Nucl. Phys. B **202**, 89 (1982).
[48] N. H. Christ, R. Friedberg, and T. D. Lee, Nucl. Phys. B **210**, 310 (1982), reprinted on p. 141 of Ref. 45.

[49] N. H. Christ, R. Friedberg, and T. D. Lee, Nucl. Phys. B **210**, 337 (1982), reprinted on p. 168 of Ref. 45.

[50] Here and elsewhere, comments about a lack of gauge invariance in tight-binding theory refer to models with intra-atomic coordinate matrix elements.

[51] J.-M. Jancu, R. Scholz, F. Beltram, and F. Bassani, Phys. Rev. B **57**, 6493 (1998).

[52] T. Hahn, ed., *International Tables for Crystallography*, vol. A (Kluwer, Dordrecht, 1992), 3rd ed.

[53] R. H. Parmenter, Phys. Rev. **100**, 573 (1955).

[54] P. Boguslawski and I. Gorczyca, Semicond. Sci. Technol. **9**, 2169 (1994).

[55] L. P. Bouckaert, R. Smoluchowski, and E. Wigner, Phys. Rev. **50**, 58 (1936).

[56] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Wiley, New York, 2000), 2nd ed.

[57] Magnetic monopoles could be included in this formalism by paying close attention to phase changes of $2\pi n$ around a closed path.[36,46,80] Such phase changes have been swept under the carpet here because magnetic monopoles are ordinarily of no interest in tight-binding theory. An alternative approach to magnetic monopoles is given in Ref. 38.

[58] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), 2nd ed.

[59] Equation (4.20) may also be derived from Eq. (4.21).

[60] G. Arfken, *Mathematical Methods for Physicists* (Academic, San Diego, 1985), 3rd ed., p. 55.

[61] J. J. Sakurai, *Modern Quantum Mechanics* (Addison-Wesley, Reading, MA, 1994), revised ed., p. 54.

[62] As noted by Graf and Vogl,[23] the relation $[x^\alpha, p^\beta] = i\hbar\delta_{\alpha\beta}$ can never be satisfied in a finite basis, because $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ in a finite basis.

[63] If the lattice happens to be a Bravais lattice, however, then one can show that

$$\sum_j \langle \mathbf{X}_i | [x^\alpha, p^\beta] | \mathbf{X}_j \rangle \Omega_j = \frac{i\hbar}{2\Omega_i} \sum_j d_{ij}^\alpha S_{ij}^\beta = i\hbar\delta_{\alpha\beta},$$

where the second equality follows from Eq. ($10\frac{1}{2}$) of Ref. 49 and the fact that a Bravais lattice has a center of inversion at the midpoint between two lattice sites,[81] so that the center of mass of the surface S_{ij} is just $\frac{1}{2}(\mathbf{x}_i + \mathbf{x}_j)$.

[64] C. H. Xu, C. Z. Wang, C. T. Chan, and K. M. Ho, J. Phys. Condens. Matter **4**, 6047 (1992).

[65] L. L. Foldy and S. A. Wouthuysen, Phys. Rev. **78**, 29 (1950).

[66] A. Messiah, *Quantum Mechanics* (North-Holland, Amsterdam, 1962).

[67] M. L. Cohen and T. K. Bergstresser, Phys. Rev. **141**, 789 (1966).

[68] P. Vogl, H. P. Hjalmarson, and J. D. Dow, J. Phys. Chem. Solids **44**, 365 (1983).

[69] J. R. Chelikowsky and M. L. Cohen, Phys. Rev. Lett. **31**, 1582 (1973).

[70] J. R. Chelikowsky and M. L. Cohen, Phys. Rev. B **10**, 5095 (1974).

[71] J. R. Chelikowsky and M. L. Cohen, Phys. Rev. B **14**, 556 (1976).

[72] F. Bassani and G. Pastori Parravicini, *Electronic States and Optical Transitions in Solids* (Pergamon, Oxford, 1975).

[73] G. Gilat and L. J. Raubenheimer, Phys. Rev. **144**, 390 (1966).

[74] L. R. Saravia and D. Brust, Phys. Rev. **171**, 916 (1968).

[75] S.-Y. Ren and W. A. Harrison, Phys. Rev. B **23**, 762 (1981).

[76] D. Brust, Phys. Rev. **134**, A1337 (1964).

[77] J. F. Cornwell, *Group Theory in Physics*, vol. 1 (Academic, San Diego, 1984), pp. 213 and 233.

[78] D. E. Aspnes and A. A. Studna, Phys. Rev. B **27**, 985 (1983).

[79] L. X. Benedict, E. L. Shirley, and R. B. Bohn, Phys. Rev. B **57**, R9385 (1998).

[80] T. A. DeGrand and D. Toussaint, Phys. Rev. D **22**, 2478 (1980).

[81] G. Burns and A. M. Glazer, *Space Groups for Solid State Scientists* (Academic, San Diego, 1990), 2nd ed., p. 53.